Serial Analysis of Genes Expressed In Normal Human Glomerular Mesangial Cells

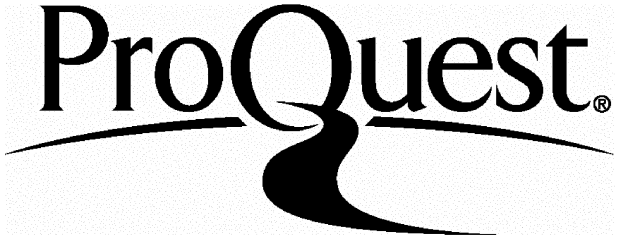
A thesis presented for the degree of Doctor of Philosophy by

James Maxwell Wilkinson

UNIVERSITY COLLEGE LONDON
University Of London

Royal Free & University College London Medical School
Department of Medicine
Centre for Nephrology
Royal Free Campus

ROWLAND HILL STREET
LONDON
NW3 2PF
UNITED KINGDOM



Abstract

Advances in sequencing based genomics like the Human Genome Mapping Project (HGMP) have meant that the majority of the estimated human genes have been at least partially sequenced. The variation in expression of a set of essentially identical genes will provide information on the molecular basis of phenotype. Serial analysis of gene expression (SAGE) is based on the ability to assign an individual transcript to a ten base pair 'tag', and the technology facilitating rapid sampling of such tags.

Glomerular mesangial cells (MC) are considered to play a major role in the development of renal disease and *in vitro* culturing of MC's has become a model system with which to study the molecular mechanisms of glomerular pathology. To this end, a SAGE project was undertaken to identify genes expressed in normal human mesangial cells (NHMC).

Primary normal human mesangial cells were cultured for periods up to 96 hrs. A total of 46,219 tags were sampled (14,953 unique tags). Tags were mapped to 20,382 sequences. Of these 79% of tags mapped to characterised cDNAs, 16% tags mapped to ESTs. 5% of tags failed to match any database entry. The most abundant tags mapped to ribosomal genes or genes associated with the cytoskeleton. Represented in the top ten tags were the matricellular genes transgelin (1.2%), SPARC (1%) type IV collagen (0.5%) and fibronectin (0.53%), which support the notion that the MC is a producer and re-modeller of the glomerular extracellular matrix (ECM). The contractile nature of MC was apparent with the high abundance of contractile proteins like myosins and tropomyosins.

Also apparent in the transcriptome were lineage specific isoforms of several genes, supporting the myoblastoid linage of MC. Comparing the transcriptomes of the MC to other libraries revealed a high correlation between cells in the same lineage as MC, such as astrocytes, smooth muscle cells and fibroblasts when compared to libraries sampled from heart, liver and various other unrelated cell lines. Understanding gene expression in the mesangial cell facilitates a greater understanding of its role in renal pathology.

Acknowledgements

The work described in this thesis was carried out in the Centre for Nephrology at the Royal Free Hospital in London and was in part supported by a grant from the British Diabetic Association (Grant No. BDA: RD98/0001854). To begin, thanks must go to Centre Director and my supervisor, Prof. Steven H Powis, who first brought my attention to the method of SAGE, suggested I look into it, then provided me with the opportunity to do so.

The other members of the Centre for Nephrology also require acknowledgement for assistance at the Royal Free Hospital, continually answering endless questions regarding the locations of various pieces of equipment and reagents and the proofing of manuscripts as they emerged. Also requiring acknowledgement are the departmental secretaries who assisted with all my office requirements.

Heartfelt thanks must also go to members of my family, friends and especially my partner, who are all as relieved as I am that this particular project is ending. I could not have wanted more support from any of them, indeed only ever received encouragement.

Finally acknowledgement must be brought to my father who has always encouraged me to follow the path I had chosen, however ill conceived, and afforded me the education that prepared me for the opportunities and choices with which I was presented. This thesis represents a fulfilment of one such path and is dedicated to him, Kenith M Wilkinson, and in memory of my mother, Agnes M Wilkinson.

Table of Contents

Figures

Tables

Equations

Abbreviations

General abbreviations used throughout this thesis. A complete list of gene abbreviations is presented in APPENDIX 1

ABBREVIATION	DESCRIPTION
× g (rcf)	Relative Centrifugal Force 'g'
βME	Beta Mercaptoethanol
β-NAD	Beta Nicotinamide Adenine Dinucleotide
A_{260}	Absorbance at 260nm
AE	Anchoring Restriction Enzyme (Nla III)
AGE	Advanced Glycation End products
Amp	Ampicillin
AMV	Avian Myeloblastosis Virus
AR	Aldose Reductase
ARI	Aldose Reductase Inhibitor
ATCC	American Type Culture Collection
BLASTn	Basic Local Alignment Search Tool nucleotide
bp	Base Pairs
BSA	Bovine Serum Albumin
BW	Binding & Washing buffer
cDNA	Complementary DNA (to mRNA)
CGAP	Cancer Genome Anatomy Project
DAG	Diacyl Glycerol
DCCT	Diabetes Control and Complications Trial
DM	Diabetes Mellitus
DMSO	Dimethyl Sulfoxide
DN	Diabetic Nephropathy
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleoside Triphosphates (dATP,dCTP,dTTP and dGTP)
dsDNA	Double Stranded DNA
DTT	Dithiothreitol
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic Acid
ESRF	End Stage Renal Failure
ET-1	Endothelin-1
EtBr	Ethidium Bromide
FBS	Foetal Bovine Serum
FCS	Foetal Calf Serum
GAPDH	Glyeraldehyde-3-Phosphate Dehydrogenase
GF200	Gene Filter release 200
GFAT	Glutamine:Fructose-6-Phosphate Amino Transferase

ABBREVIATION	DESCRIPTION
GSH/GSSH	Glutathione (Reduced/Oxidised)
GTE	Glucose Tris EDTA buffer
GTT	Gene-To-Tag
HBSS	Hanks Balanced Salt Solution
HGMP	Human Genome Mapping Project
HMCL	Human Mesangial Cell Line
IAC	Chloroform: Isoamyl Alcohol (24:1)
IDDM	Insulin Dependent Diabetes Mellitus
IMAGE	Integrated Molecular Analysis of Genome Expression
kb	Kilo basepairs
LB	Luria Broth
LPS	Lipo-polysaccaride
MC	Mesangial Cell
mECM	Mesangial Extracellular Matrix
MGO	Methyl glyoxyl
MMV	Moloney Murine Leukemia Virus
mol	Amount of Substance (SI)
MOPS	(3-(N-Morpholino)propanesulfonic acid
mRNA	Messenger RNA
MW	Molecular Weight
NCBI	National Centre for BioInformatics
NHMC	Normal Human Mesangial Cells
NIDDM	Non Insulin Dependent Diabetes Mellitus
NOS	Nitric Oxide Synthase
oligos	Oligonucleotides
P/IAC	Phenol:Chloroform:Isoamyl Alcohol (25:24:1)
PAGE	Polyacrylamide Gel Electrophoriesis
PAI-1	Plasminogen Activator Inhibitor 1
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDGF	Platlet Dervied Growth Factor
PEG_{8000}	Polyethylene Glycol (average MW 8000)
PKC	Protein Kinase C
PMA	Phorbol 12-Myristate 13-Acetate
$PolyA^+$ RNA	RNA with poly A tails
PPP	Pentose Phosphate Pathway
R/T	Room Temperature
RAGE	Receptor for AGE's

ABBREVIATION	DESCRIPTION
rcf	Relative Centrifugal Force
RNA	Ribonucleic Acid
ROS	Reactive Oxygen Species
rpm	Revolutions Per Minute
RT-PCR	Reverse Transcriptase PCR
rtRT-PCR	Real-Time RT-PCR
SAGE	Serial Analysis of Gene Expression
SAGEmap	Data files relating SAGE tags to Unigene Clusters
SDS	Sodium Dodecyl Sulphate
SNP	Single Nucleotide Polymorphism
SSC	Saline Sodium Citrate (0.015M Citrate,0.15M NaCl)
STZ	Streptozotocin
ssDNA	Single Stranded DNA
SSPE	Saline Sodium Phosphate EDTA (10mM Phosphate,150mM NaCl, 1mM EDTA)
SYBR Green 1	Syanogen Bromide green 1
$T_{10}E_1$	Tris EDTA buffer (10mM Tris, 1mM EDTA, pH8.0)
TAE	Tris Acetate EDTA buffer
TAPS	Tris-Acetate PCR Buffer (Qiagen)
TBE	Tris Borate EDTA buffer
TE	Tagging Enzyme (Bsm FI)
TGFβ	Transforming growth Factor beta
THP-1	Human Monocyte Cell Line
TRNA	Total cellular RNA
TTG	Tag-To-Gene
UTR	Untranslated Region
VEGF	Vascular Endothelial Growth Factor

CHAPTER 1

1 INDRODUCTION

1.1 THE HUMAN GENOME MAPPING PROJECT (HGMP) & FUNCTIONAL GENOMICS

Sequencing the entire human genome will provide a complete catalogue of genes. Such a catalogue will not only contain information on the name and sequence of each gene, but also contain many variations and mutations, and as such can be thought of as the biological equivalent to a chemical periodic table (Fields et al., '94, Lander, '96, Fields, '97). This catalogue of all genes required to define a living organism will also contain a basis of classification for these genes. From its inception in the early 1980's the task of sequencing the entire human genome was considered a methodical sequencing project that would be completed in 20-25 years, and was designed to reveal three levels of genomic classification, reflected in the three analytical layers of the sequencing project. The first step produced the genetic maps of gene units, the second mapped the physical location of these gene units and the final step would be completing the DNA sequence of the genome (Lander et al., '01).

Together with this methodical approach came the unexpectedly efficient method of random sequencing of cDNA in conjunction with an alternative experimental procedure, the whole genome shotgun (WGS). This process complemented the time consuming sequencing of hierarchical contigs, concentrating rather on the high through-put single read sequencing of expressed genes easily obtained from cloned cDNA and genomic fragments that were produced directly from genomic DNA (Venter et al., '01). These EST (Expressed Sequence Tag) libraries provided a resource for molecular and cellular biologists to use in the ever-expanding search for the functional significance of new genes. This was a useful resource for the dissemination of coding sequences within the genomic DNA, which in vertebrates is remarkably diffuse with some genes reaching into hundreds of kilobases for a mRNA that is only 1.5kb. The result was a draft map created by two alternative methods, one of which was methodical and rational, the other efficient and resourceful.

The progress of the human genome project has been remarkably swift, mainly due to technical advances in sequencing efficiency and computational power. The published draft map claimed to cover 96% of the euchromatic genome. Together with

the already available public data the level of coverage is currently believed to be 94%. The sequence therefore is not complete and still in the initial draft form. Many gaps and ambiguous regions exist within the current data set, and these require resolving, as does the remaining 4% of genomic sequence. Despite this, the HGP will form a valuable archive of data and a resource for future study.

In its current state, the HGP has facilitated discussion in a very general sense. Questions such as the distribution of genes and repeated elements along the chromosomes, the relationships to homologous genes in other species and the allelic differences in populations have all been given insight through the HGP. Additionally, the identification of widely and evenly dispersed single nucleotide polymorphisms (SNPs) mean even individuals within a population may be classified (Sachidanandam et al., '01, Stoneking, '01). However, detailed questions regarding gene expression have been revealed to be more complicated than once thought.

1.1.1 ESTIMATING THE COMPLEMENT OF GENES IN THE GENOME

The coding capacity of the genome is an ongoing topic of debate. The window of estimation as the HGP progressed was between 30,000 and 150,000 genes (Lander, '96, Deloukas et al., '98). Following the initial draft of the genome this has now been downgraded and estimates now run to 30,000-40,000 genes (McPherson et al., '01). To date the HGP has described location and sequence for 22,000 of these genes and other mapping projects claim to have data to place 26,000 genes although proprietary licence precludes easy non-commercial access (Claverie, '01).

The human genome is around 30 times larger than that of the fly and nematode and around 250 times larger than the genome of yeast. This disparity in overall size is counterintuitive as humans only possess two to three times the number of genes than fly and nematode genomes and these are present in only 3-4% of the DNA sequence (Rubin et al., '00). A large amount of human DNA (46%) is repeated motifs that possess no known function, a phenomenon common in vertebrates. The remaining DNA is composed primarily of promoter sequences, transcriptional elements, introns and sequence of unknown function.

While these estimations are neither static nor complete, it is becoming apparent that the number of genes is only one mechanism for assessing the genome. The literature is rich with reports of alternative transcripts for the same gene, often the result of multiple transcription start sites, alternative splicing of introns, premature or delayed termination, and transcription artefacts (Montoliu et al., '90, Lin et al., '93, Ayoubi and Van De Ven, '96, Rogaev et al., '97). Indeed the complex nature of the transcriptional complement of the cell is currently an area of intense research and forms the primary experimental approach of this thesis. The HGP has supported the notion of multiple transcripts with data suggesting numerous ESTs clusters over the same gene. Recently, a report used 700,000 ESTs and assembled them into 15,000 full-length mRNA clusters (Camargo et al., '01). These authors estimated that over 80% of human genes are now at least partially sequenced and enough ESTs have been compiled to facilitate the building of a scaffold over a gene which would experimentally close many of the gaps present in the coding regions of the HGP draft sequence. However, this scaffold revealed discrete clustering of ESTs over genes, which suggests transcriptional units rather than discrete genes.

With the advances in sequencing based genomics and public and private EST libraries, the majority of the estimated human genes have been at least partially sequenced. The next major biological challenge will be to assign a functional significance to this genetic information, so called functional genomics (Deloukas et al., '98). The hypothesis that the phenotype of a cell is essentially defined by the genes it expresses forms the basis of functional genomics. The dynamic link between the information contained in the cell's genome and its phenotype provides an opportunity to test this hypothesis and investigate the transcriptional elements associated with changing phenotype.

1.2 THE THREE MOLECULAR LEVELS OF A CELL

While the HGP is without doubt a key achievement in biological science, the genome is essentially a static entity that contains information much in the same way as a dictionary contains a language or a periodic table contains the chemical elements. Understanding the order and location of genes within a genome represents one of

several levels of genomic expression. The expression of a genome is an important and natural extension to the HGP and studies of gene expression are numerous in the literature. Somewhat ignored during the early intensive activity of the HGP is the study of the proteins encoded in the genome. A gene will usually produce a protein (or other non-peptide catalytic unit such as ribozymes), and it is the timely and ordered functioning of the protein, which forms the metabolic dimension to the cell.

Therefore, the human genome project represents only the first dimension of the molecular basis of a cell. Beyond the genome is the elucidation of the mechanisms of gene transcription and the entire transcriptional profile of a cell, the transcriptome. The final mechanism will be the complete complement of proteins and their specific roles in the biochemistry of the cell, the proteome. These increasing levels of resolution reflect increasing levels of complexity. The genome may be complicated in the sense that is a large amount of data, but it is essentially static and thus more readily quantifiable. The transcriptome is a dynamic entity and changes in response to stimuli, differentiation lineage and cellular life cycle. A transcriptome may be considered unique to a cell or cell type within a mass of thousands that constitute the organism. The proteome is by far the most complicated of the three. Not only does the sequence of a protein indicate its function, the same protein can be procured, translocated, structurally altered, phosphorylated, glycosylated, secreted and sequestered, all in a specific manner. The functional significance of this 'protean' expression, from gene to transcript to functional catalytic unit, creates a truly complex set of data, and increasing complexity leads to greater inaccuracy in measuring the relationships between these levels.

1.2.1 THE GENOME

The sequential combination of four nucleotide base pairs over some 2.91 Gbp has revealed more complex elements than merely a series of protein coding regions (Deloukas et al., '98, Ferea and Brown, '99, Caron et al., '01, McKusick, '97). The genome appears to show many variations in the distribution of genes, mobile elements (transposons), GC rich regions, CpG islands and areas of linkage disequilibrium. There appear to be some 30,000-40,000 protein coding regions, as estimated from the two major mapping projects and a heritage that reaches across many species barriers, with gene homologues from bacteria, fungi and plants. Segmental duplication from other regions of the human genome appears more common that other genomes like those of

Drosophila melanogaster and *Caenorhabdtitis elagans*. Over 1.4 million single nucleotide polymorphisms (SNP) have been identified and their frequent occurrence across the genome should facilitate further mapping of linkage disequilibrium of genes throughout the population (Sachidanandam et al., '01).

1.2.2 THE TRANSCRIPTOME

The second level of complexity is the controlled transcription of the genome within the organism. Transcription of genes from DNA into mRNA is a fundamental factor in gene expression. In the context of the human body, each of the several thousand cell types is believed to have a series of unique patterns of gene expression designed for specific biological function at particular times in their cycle. External and internal factors can modulate the levels of gene expression and lead to altered phenotype associated with physiological differentiation and disease. This variation in gene expression between cell types provides us with the next level of genetic complexity.

Researches have long recognised the relationship between a gene's expression and its functional role. This reasoning has lead to the initiation and continuing search for differentially transcribed genes in disease states and more simply in altered phenotype. The variation in expression of a set of essentially identical genes has provided biology with clues to the functional roles of genes in the cellular context as well as a molecular basis for phenotype. Technical advances have lead to several methods for investigating genome-wide expression. Many are based largely on the partial gene sequences derived from HGP and EST libraries. With the development of high resolution fluorescent detection and solid support for hybridisation targets it is now possible to monitor the level of expression of tens of thousands of genes on a single microscope slide (reviewed in (Ferea and Brown, '99). Alternatively, one may enrich for differentially expressed genes by selective amplification using anchored redundant sets of PCR primers (Liang and Pardee, '92).

Combining technical advances in transcript monitoring with the HGP data has resulted in 'mining' for genes previously uncharacterised in model systems. This method is not strictly hypothesis driven and as such creates profiles of gene expression based on empirical observations and not preconceived models. This experimental

approach is useful in two ways. First, it allows the simultaneous monitoring of many genes within a single experiment and secondly, it has facilitated the annotation of genes not previously associated with the particular model system by virtue of the ability to include them in an analysis. This was evident in the study of fibroblasts and their proliferative response to serum, an *in vitro* model for the wound healing process. A total of 9000 genes were monitored over several time scales but because the genes present on the array were not limited to those known to be involved in proliferation, a complex pattern relating to the physiology of wound healing was unexpectedly revealed (Iyer et al., '99).

More recently, the clustering of EST data to chromosomal maps has revealed regions of chromosomes that contain differing densities of gene transcription, with some regions particularly rich for highly transcribed genes. Additionally, differences between the transcriptional potential of the chromosomes became apparent. Chromosome 19 displayed an extremely high level of transcription compared to other chromosomes, while chromosome 13 displayed very little transcriptional activity at all. Compiling these maps has revealed novel genes involved in disease (Caron et al., '01).

1.2.3 THE PROTEOME

The function of most genes will inevitably lead to the study of the proteins they code (Blackstock and Weir, '99, Malakoff and Service, '01, Fields, '01). The progression from DNA sequence to transcriptional intermediates and translation and modification into functional protein regards the full picture of gene expression. Proteomics has often been neglected with the advent of the HGP but ultimately it will rise to the forefront of biological research as the fruits of the HGP are realised. The current dilemma in proteomics is similar in fashion to that facing the dynamic transcriptome and opposed to the essentially static genome. This dilemma is the jump from the study of a single gene product and its membership in a specific pathway to the study of many possibly thousands of gene products and their intermingling in the seemingly chaotic biochemistry of the cell. While this has been partially addressed in the transcriptome with improvements in transcription analysis, proteomics has not enjoyed such advancement.

The knowledge of protein active sites is only a step in the understanding of the protein function. A protein may have many active sites and participate in a single biochemical pathway or only one active site and possess activity in many pathways, as is evident in the complexities of secondary messengers and signal transduction. Proteins can act alone or in concert with many others. The cellular location of a protein can also affect a control mechanism as with PKC isoforms, only aligning substrates with catalysts at certain times of the cellular cycle.

One particularly useful analysis is determining the total number of distinct protein units in an organism, the 'core proteome' (Schuler et al., '96, Ferea and Brown, '99, Mushegian, '99, Ison et al., '00, Persson, '00). While grouping protein paralogs as single units is a crude indicator of complexity, the comparison of so assembled core proteomes has revealed another counterintuitive phenomenon hinted when comparing the genomes. It appears that despite the clear magnitudes of difference between the single-cell yeast, *Saccharomyces cerevisiae* and the metazoan fly, *D.meloanogaster*, there is only a two-fold difference in the core proteome (Rubin et al., '00). When this is added to the hypothesis that the large human genome is a result of multiple duplications, this leads to the idea that there are probably not many more protein members in the human 'core proteome' than there are in that of the fly or nematode.

The most widely studied protein regulation has been the phosphorylation states of the protein unit, with the addition of or cleavage from inorganic phosphate responsible for the catalytic action of the protein. Similarly, the action of glycosylation has been proposed to be another control mechanism analogous to phosphorylation, an example being the increased binding of glycosylated transcription factors to promoter regions compared to their un-glycosylated relatives (Hart, '97). Genome and transcriptome analysis will only confront the proteome at its basic level, i.e. switching genes on and off, and the post-transcriptional processing to include or exclude particular exons. At its best, this will only provide clues to the proteins that are involved in invoked cellular processes. What will follow will be the branching of this genome data into functional investigations.

1.3 A DYNAMIC LINK BETWEEN GENOTYPE AND PHENOTYPE

With the near completion of the human genome mapping project, as well as other genome projects of the mouse, nematode, numerous viruses, various plants and invertebrates, a large dataset has been generated and the first step to a greater understanding of the dynamic functioning of the living organism has been achieved (Velculescu et al., '00, Caron et al., '01). The genome represents a structural basis for what is essentially the plan of cellular function. As with most plans its mere presence is not in itself informative of an entire picture but a guide to the arrangement of raw products into functional molecules. The central dogma of molecular biology is that genes in DNA contain the information required for the synthesis of proteins from a relatively small set of nucleotides and amino acid building blocks. The cellular proteins affect their function through catalytic or specific sites before being degraded back to amino acids. Protein turnover is constant through the life of the cell. From this it appears that DNA in itself is a relatively benign cellular molecule affecting no real function except as a 'user manual'. However, the information contained in DNA is more complex than simply a code for protein synthesis. Continued study of DNA has revealed that there are controlling mechanisms encoded within the DNA molecule that affect the transcription of genes, the processing of pro-transcripts and even the signalling of specific translation and processing of early proteins.

The variation in a gene expression is generally a much richer source of information than allelic variation in a sequence. The variation of global gene expression underlies the mechanisms for variation of phenotype within multi-cellular organisms. This is rationalised from the notion that the expression of a set of essentially identical genes, and more particularly the variation of expression of those genes, as the fundamental basis for phenotype. Studying the variation of gene expression requires two levels of data accumulation. Primarily, differential gene expression is used as a tool to test cause and effect hypotheses, i.e. the association of the increased or decreased gene expression with actively altered phenotype. But in a more general setting the collection of global gene expression data permits investigation of the molecular basis of phenotypic variation among cells and individuals.

1.4 THE STUDY OF TRANSCRIPTOMES

Several techniques have been developed which use changes in gene expression to identify and characterise novel genes, or are designed specifically to search for such genes. An understanding of the available technologies and their advantages and disadvantages will benefit the understanding of transcription analysis and provide a basis for the inclusion or exclusion of a specific technique. A brief review of techniques also provides an understanding of the data types that are generated.

1.4.1 DIFFERENTIAL DISPLAY

Differential display was first described as a tool to identify genes that were differentially regulated by their selective amplification using anchored and random primers in PCR reactions (Liang and Pardee, '92). Products of various sizes were enriched as the genes were 'turned on', then resolved on polyacrylamide gels where the induction or suppression of a gene was seen as an appearance or disappearance of product bands. Though initially useful and to a degree reproducible, differential display tended to produce a large amount of false positives and was reliant on redundant, anchored oligos, that showed non-random annealing and a high contamination rate in the amplification stage (Matz and Lukyanov, '98, Ledakis et al., '98, Frost and Guggenheim, '99). The next generation of techniques based on differential display, e.g. suppressive-subtractive hybridisation, have been used to study the effects of high glucose on cells of renal origin and has successfully identified a number of up-regulated genes. These include CTGF, Thrombospondin-1, a zinc finger homologue and the amiloride sensitive sodium channel as well as a number of unidentified transcripts (Page et al., '97, Holmes et al., '97, Death et al., '99).

1.4.2 MICRO ARRAYS AND GENE CHIPS

Micro-arrays and the so-called gene chips are essentially high density dot blots where experimental mRNA is labelled with a tracer molecule, generally fluorescent dyes or radio-labelled nucleotides, and hybridised to immobilised gene fragments in simple hybridisation experiments (see FIGURE 1.1). In conventional dot blots these gene fragments would number 5-50 and could be manually constructed in the laboratory (Bernard et al., '96). With micro-arrays thousands of genes are robotically arrayed and

bound on a nylon filter and with gene chips tens of thousands of genes are bound on glass slides (Schena et al., '95). Both these methods facilitate the simultaneous monitoring of gene transcripts based on homology and represent the most popular and accessible method of global transcription analysis.

Micro-array and gene chip technology has been used to examine the pathology of many disorders, including diabetes mellitus (DM). (Wada et al., '01). Despite the apparent widespread use of this technology and its application in global expression analysis there are currently far more reviews describing experiments and possible uses than original research publications. No doubt this will change as the technology advances though there certainly appears problems associated with interpretation of data, especially the inability to standardise platforms across the research community (Ermolaeva et al., '98, Boguski, '99, Scherf et al., '00).

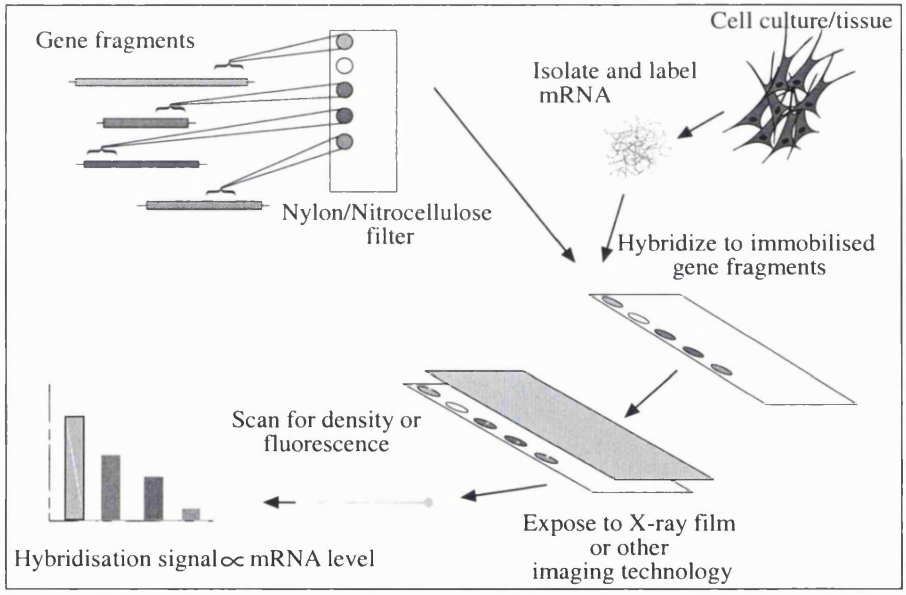


FIGURE 1.1. FLOW CHART ILLUSTRATING THE STEPS INVOLVED IN HYBRIDISATION.

Briefly, gene fragments (targets) are immobilised on solid supports, generally nylon membrane but currently glass slides are also used. The capacity of each filter/slide is increasing and the most recent glass slide can accommodate more than 20,000 targets on a slide $2cm^2$. mRNA from an experimental sample is isolated and labelled with a radioactive or other tracer marker (query). The labelled query is hybridised to the immobilised target and level of hybridisation is proportional to the intensity of the signal, which is proportional to the abundance of query in the experimental sample. Comparing two or more hybridisation signals generated from different queries reveals genes of differing abundance

1.4.3 IN SILICO MINING

In silico mining is a relatively new term in molecular biology. The literal meaning regards the exploration of genomic, proteomic and transcriptional data within the setting of data storage and retrieval i.e. bioinformatics. With the advent of molecular techniques in DNA cloning, sequencing and expression, a vast repository of data has amassed, initially nucleic acid and protein sequence (see TABLE 1.1). While the structuring of this data and its retrieval are uppermost in its usefulness, the notion of hypothesis driven research is beginning to be supplanted by *in silico* mining in attempting to access the information contained in these vast datasets. A primary and current use of *in silico* mining is the automated searching for coding regions (CDS) and open reading frames (ORFs) in raw genomic sequence data. Such regions are then annotated as containing putative genes and transcriptions sites within the genome.

(Rana et al., '01) presented an example of *in silico* mining. Using only homology searches the gene structures of adenylate cyclases were determined. Adenlyate cyclases are a multigene family consisting of some 9 transmembrane isoforms, ADCY1-9, which catalyse the formation of cAMP from ATP. The cloning of the DNA for this gene family was troublesome because, as inferred from other species, the genes for ADCY contain numerous large introns, which complicate most standard cloning techniques, including PCR and recombinant based technologies. Alignments of homologous sequences using BLASTn at NCBI revealed numerous contigs that were unassigned and unaligned in the HGMP. Further, when fully analysed they revealed the 21 exons of the ADCY gene and confirmed that indeed it was spread over 18.4kb. While such data may have been determined from a cloning project the *in silico* experimental approach, or data mining, was conducted with minimum laboratory time and only required access to relatively inexpensive computing equipment and basic knowledge of nucleic acid bioinformatics.

Such *in silico* mining has also been used in more complex analysis of whole EST collections. The mining of some 4 million ESTs using a four-step procedure resulted in the identification of some 152 potentially differentially regulated genes in invasive ductal breast tumours (Schmitt et al., '99). This analysis was carried out with only access to computing power and vast EST collections. Clearly this *in silico* approach represents an important tool for the experimentalists of the genome and

validates the continued efforts in EST library construction and dissemination of data (see TABLE 1.1).

1.4.4 SERIAL ANALYSIS OF GENE EXPRESSION

The techniques described above have inherent limitations. Differential display is highly dependent upon the sensitivity of degenerate primers and may identify only a fraction of candidate genes, while micro arrays, gene chips and *in silico* mining still require the use of known cDNA sequences (Heller et al., '97, Ledakis et al., '98, Frost and Guggenheim, '99, Rana et al., '01). The generation of large EST libraries has proved useful in the analysis of the primary DNA data emerging from the HGMP, but the task of randomly sequencing entire cDNA libraries is laborious, even using current automotive standards, and represents a large investment in time and resources.

A relatively new method of transcriptional analysis has been described, serial analysis of gene expression (SAGE), which addresses many of these problems (Velculescu et al., '95). SAGE is based on the mathematical calculation that the genetic code of a 9bp fragment contains sufficient information to discriminate between 4^9(262,144) individuals. As this represents a several fold redundancy of the human genome it should be possible to identify any gene transcript from a 9bp sequence. Furthermore, the use of short tag sequences permits their serial incorporation into vectors to facilitate high throughput sequencing (see FIGURE 1.2).

A typical SAGE analysis would require that mRNA be isolated from the cell or tissue under analysis and the cDNA synthesised from this mRNA processed in such a way that serial concatemers of 20-30 tags are isolated and sequenced in a single sequencing reaction. Detection of unique tags will be directly proportional to the abundance of the mRNA they represent. SAGE tags are identified or mapped back to their transcripts by searching public access databases like those in the NCBI (www.ncbi.nlm.nih.gov) and candidate genes are investigated using more standard techniques. The result is a high-resolution map of the transcriptional elements within the cell or tissue studied. This provides a library of genes and their transcriptional frequencies within a particular cell at a particular time.

To date SAGE has been used to successfully study genes expressed in many human cell systems, both *in vitro* and *in vivo*, and has amassed some 2.2 million tags, representing 454,836 unique tags, from some 136 SAGE libraries. SAGE data from *Arabidopsis thaliana*, *Mus musculus* and *Rattus norvegicus* libraries are also collected. (http://www.ncbi.nlm.nih.gov/SAGE) (see TABLE 1.1)

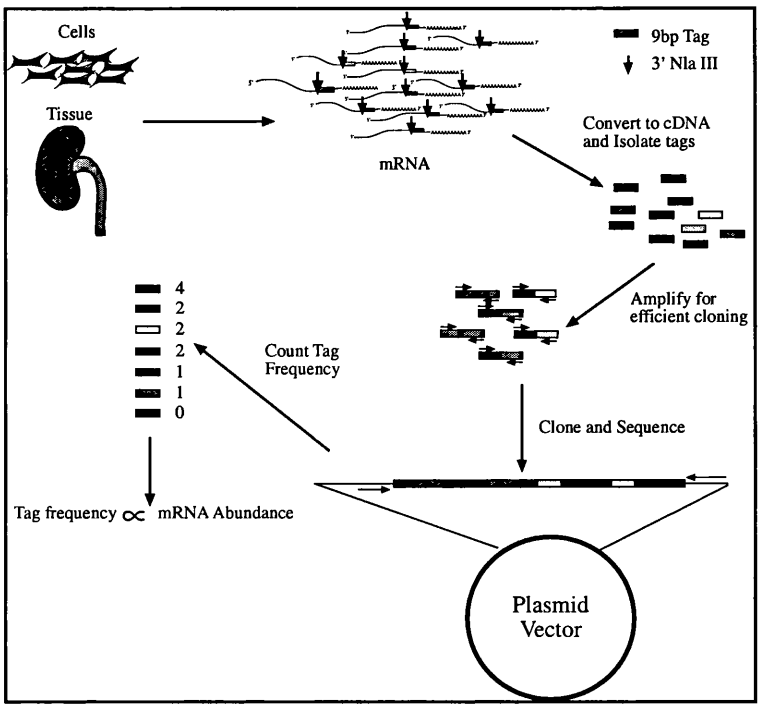


FIGURE 1.2. SCHEMA OF THE STEPS INVOLVED IN A SAGE ANALYSIS.
mRNA is isolated from experimental samples and converted to cDNA. Tags 9-10bp are isolated from points in the cDNA defined by the anchoring restriction enzyme *Nla* III (AE). These tags are amplified, concatenated, cloned and sampled by sequencing. The frequency of each tag is proportional to the abundance of the gene it represents. Using this protocol for samples under different experimental conditions allows the identification of genes that are differentially transcribed in response to the experimental stress.

Information Web Sites	Platform	Comments
www.ncbi.nlm.nih.gov/SAGE	Serial Analysis of Gene Expression	Database for the entire SAGE data including mapping and digital Northern data.
www.ncbi.nlm.nih.gov/UniGene	cDNA clustering	Creates (builds) gene clusters from submitted gene sequence and EST data.
www.affymetrix.com	Micro array	Current market leader in high density gene 'chips' and scanning technology
www.geneindex.org	Micro array	Microchip data profiling gene expression data in many cell systems
www.nhgri.nih.gov/DIR/LCG/15K	Micro array	Database for warehousing micro array data together with profiling and mining algorithms
www.ncbi.nlm.nih.gov/CGAP	Genome annotation	Warehouse for cancer genome anatomy project
www.mgri.org	Genome annotation	Warehouse for genome annotation
www.sagenet.org	SAGE	Site providing SAGE technology, references and links to warehouse data.

TABLE 1.1. INTERNET SITES PROVIDING BIO-INFORMATICS AND DATA WAREHOUSING.

The dissemination of genomic data and collaboration is fundamental to the early days of the HGP. This table represents a small selection of web sites providing data warehouses and technology, some of which are used in this thesis. Currently there are two major technologies involved in genome annotation, serial analysis of gene expression (SAGE) and micro-array data. Both are well represented on the Internet and while micro-array data is more common the handling of SAGE data allows more straightforward sharing.

With the problems of data storage, retrieval and cross platform access faced by the micro-array and gene chip technology, SAGE offers the advantages of both warehousing and *in silico* mining. The output from a SAGE library is essentially four fields containing the tag sequence, the frequency in the library, the mapping information and some unique tag identifier. In this regard, the SAGE libraries are much simpler than their hybridisation-derived cousins and thus retrieval and cross-library comparisons are also straightforward (Adams, '96, Zhang et al., '97, Velculescu et al., '00). Digital data is based on absolute frequency within a population sample and not on a relatively arbitrary fluorescence or radioactive signal. Many believe the SAGE library is a truer picture of a transcriptional profile than any other current form of array technology, and that assignations of absolute abundance rather than relative abundance are valid. An added advantage of a SAGE library is that library membership is based on abundance and is not dependent on predetermined sequence homology. In this respect the SAGE analysis represents the next generation of bioinformatics, where discovery in based on the presence of an individual rather than an active search for that individual.

1.5 USING SAGE TO INVESTIGATE THE TRANSCRIPTION OF GENES

A further use for SAGE libraries is the application to differential analysis. Clearly if a SAGE library can accurately describe a cellular phenotype, comparison of SAGE libraries constructed in parallel in the same cells, one controlled the other under experimental conditions, will reveal tags and thus genes that are differentially transcribed in the experimental system. The nature of sampled SAGE data is essentially digital and facilitates comparison. Assigning significance to such comparison is more complicated than usual standard statistical techniques, but has been addressed in many SAGE publications and is currently used in the SAGEmap database (NCBI) (Chen et al., '98a, Lash et al., '00). A summary of the statistical methods used is presented in CHAPTER 2. Once a significance factor is used to filter the differential data, straightforward molecular techniques can be applied to verify the data.

Several key original papers highlight the primary use of SAGE as an analysis tool. The original paper described the technique and compared SAGE abundance of several genes with cDNA screening and found them to be represented at a similar level. This

was a small SAGE library with a low level of sampling (approximately1000 tags) and described the normal pancreatic profile (Velculescu et al., '95). The second important publication involved the larger application of SAGE to the yeast cell cycle with the express purpose of identifying new genes involved in the yeast cell cycle, together with estimating the complete transcriptome of the yeast cell (Velculescu et al., '97). In these experiments, more than 60,000 tags were sampled, representing 4,500 genes. Over 1,980 genes had been previously characterised while the remaining 2,520 were not. The last of the three original papers involved the large scale sampling of SAGE tags from several cancer cells, primary cells and dissected cancers, amassing some 300,000 tags that were representative of some 45,000 transcripts (Zhang et al., '97). This study identified many genes that had not previously been described in these particular cancers, but also identified the differential transcription of 500 genes occurring in cancer cells, indicating possible causative mechanisms and thus diagnostic markers or therapeutic targets.

The expanding database for SAGE tags has increased dramatically in the last few years and the ease with which data can be stored, manipulated and analysed is reflected in the usefulness of the SAGE mapping project, SAGEmap (ncbi.nlm.nih.gov/sage). SAGEmap has compiled the data from all submitted libraries and mapped the tags back to the transcripts they represent, in this case an individual EST and/or mRNA as well as the UniGene cluster they belong to (Lash et al., '00). Original publications using SAGE as a primary technique are increasing rapidly. Identifying novel disease candidates on SAGE data are also increasing, indicating a powerful tool in molecular analysis. It is likely that investment in SAGE analysis will yield large amounts of data regarding gene expression.

1.5.1 ANALYSING TRANSCRIPTOMES

The analysis of a transcriptome facilitates observing the dynamic expression of the static genome. Though a relatively immature technology, much information can be produced from the various forms of transcriptome analysis currently available. Expressed sequence tag (EST) libraries contain large amounts of information regarding the processed sequence of genes, and facilitates the clustering of transcriptional products that can be anchored onto the genome. Such libraries require relatively large investments in time and resources that may be beyond the reach of smaller laboratories,

a situation which is at least partially addressed with public access to the databases and software for *in silico* mining. Micro-arrays provide convenient access to transcriptome analysis and, with the ease of introducing a time dimension, can accelerate a truer dynamic profile of gene transcription. Micro-array construction requires access to expensive robotic infrastructure and is ultimately dependent on hybridisation kinetics and predetermined sequence. Additionally, micro-array data requires some sort of 'industry standard' so that data can be shared more easily across the various platforms available.

SAGE offers possibly the highest resolution of transcription profiling to date. The techniques are straightforward, although complicated, and data is simple and easily manipulated. The storage and retrieval of SAGE data is particularly suited to *in silico* mining and transfer between laboratories, models and organisms remarkably easy. Based on transcript abundance, the discovery of new transcripts will compete with redundant sampling in a SAGE library. Abundant genes are more likely to be well characterised, a phenomena gleaned from the early mapping projects, but are also more likely to be sampled, thus redundant in this setting. The more obscure genes or gene products may have unique tags that can discriminate between their family members, but they may be present at such a low level that large sampling projects will be required before accurate identification and quantitation is achieved.

SAGE libraries do, however, provide a more complete profile of the transcribing genome and contain information on several levels. Primary SAGE information regards the distribution of genes or gene isoforms across cells or tissue. This data is useful in areas such as differentiation and susceptibility, based on the presence or absence of a particular gene product. An example of this would be the presence of α-actin isoforms in smooth muscle cells and not in neural cells, or the presence of oxytocin receptors on cells from the mammary glands and not on fibroblasts. Perhaps the strongest application of SAGE is the discovery of genes differentially transcribed in experimental systems. SAGE does not require candidate genes to be assumed, tested or hypothesised, only that a phenotype changes and the transcripts can be sampled efficiently. Comparing the transcriptional profiles will thus have the potential to identify new uncharacterised genes based on their presence and with no other criterion.

1.5.2 SUMMARY OF TRANSCRIPTOME ANALYSIS

Genetics continues to be faced with the problems of bridging the gap between genotype and phenotype. Transcriptome analysis provides the first step in addressing this gap, and forms the basis with which to continue addressing the expression of genes from the genome through the transcriptome and proteome. Access to the data of the HGMP coupled with evolution of transcriptional analysis offers a powerful resource in understanding the transcription of gene products and any changes in this transcription that occur from experimental stimuli or disease models. The ordered and timely expression of genes represents an important level of complexity in the understanding of an organism. The disruption of such balance, and the ability to measure changes, facilitates an understanding of pathology, which could identify causal relationships and therapeutic targets.

1.6 THE PATHOLOGY AND PROGRESSION OF DIABETES MELLITUS IN TARGET ORGANS AND TISSUES

Prior to the introduction of insulin as a treatment for insulin dependent diabetes mellitus (IDDM), most patients did not survive long enough for clinical complications to develop. It was only after the introduction of insulin treatment the manifestations of diabetes (retinopathy, neuropathy and diabetic nephropathy (DN)) became serious clinical issues. Because of the tight association between insulin and glycaemic control the hypothesis that hyperglycaemia is a causative agent for diabetic pathology was proposed. The glucose hypothesis postulates that hyperglycaemia causes diabetic complications and that correction of hyperglycaemia will prevent them. Rigorous testing of this hypothesis in experimental animals demonstrated a strong correlation between elevated blood glucose and multi-organ and tissue pathology similar to those present in humans with longstanding diabetes (Pirart et al., '78, Klein et al., '88, Chase et al., '89).

The Diabetes Control and Complications Trial (DCCT) was a landmark, multi-centre randomised clinical trial designed to assess the impact of intensive therapy compared to standard therapy on the development of micro-vascular and neuropathic

complications. The primary question addressed whether intensive insulin therapy could prevent diabetic retinopathy in patients with no retinopathy. A second question was whether intensive therapy could slow the progression of early retinopathy (DCCT and Group, '93). The study found that intensive insulin therapy reduced the incidence of diabetic retinopathy by 50% after five years and continued to decrease with time. The intensive therapy also reduced the risk of progression of diabetic retinopathy by 54%.

In summary, the DCCT found that although not able to reverse the complications of IDDM, intensive insulin treatment of IDDM reduced the incidence and delays progression of diabetic pathology in multiple organ systems. Hyperglycaemia remains the hallmark of diabetes and the hypothesis persists that good glycaemic control reduces the risk for development and progression of diabetes specific pathology within the retina, peripheral nerve, vasculature and glomerular apparatus.

1.6.1 DIABETES AND THE EYE

Diabetic retinopathy is the leading cause of visual impairment in the developed world (Palmberg et al., '81, Frank et al., '82, DCCT, '95). In the eye, retinal terminal capillary damage, micro-aneurysms, leads to leaking erythrocytes and dot and blot haemorrhages. The retinal vessels are also abnormally permeable and leak serous fluid that will eventually form hard exudates. With increasing duration of diabetes the retinal vessels can become occluded and lead to ischemic infarctions in the retinal nerve layer. The response to this ischemia is the formation of new blood vessels (neo-vascularisation) and proliferation out of the retinal surface and into the vitreous cavity. The new vessels are fragile and tend to bleed into the vitreous cavity resulting in vision obstruction. These vitreous haemorrhages will be resorbed but the fibro-proliferative changes that ensue will result in retinal traction, eventual detachment and loss of vision.

1.6.2 DIABETES AND THE NERVOUS SYSTEM

Diabetic neuropathy can take many clinical forms and is generally due to peripheral nerve segmental demyelination and changes in the capillaries supplying the nerve tissue leading to sensorimotor and autonomic dysfunction (DCCT and Group, '93). The greatest risk is the manifestation of peripheral neuropathy as foot trauma and diabetic ulcers, although neuropathy can also affect the gastrointestinal motility, erectile

function, bladder function, cardiac function and vascular tone (Nathan, '92). Not only do the nerve fibres degenerate in DM but regeneration mechanisms are short lived and fail to progress. This failure of neuronal buds to mature and progress creates the progressive neuropathy, and is currently believed to be a result of three mechanisms; metabolic dysfunction in the neurone, ischemic effects caused by vasculature abnormalities and deleterious effects of protein glycation on the supporting Schwann cells and ECM (King, '01, Cameron et al., '01).

1.6.3 DIABETES AND THE VASCULATURE

Vascular impairment is also a chronic complication of diabetes with accelerated atheroma resulting in premature, aggressive coronary artery, cerebrovascular and peripheral vascular disease. Endothelial dysfunction appears to be a common starting point for diabetic vascular disease with subsequent involvement from other cells of the vasculature, predominantly smooth muscle cells. Early haemodynamic changes include an increase in blood flow in the skin, retina and glomerulus. This increase in blood flow is seen before structural changes and is reversible early in DM. Once progressed however, the structural alterations become irreversible. Increases in blood flow and micro-vascular pressure cause leaking and thickening of the capillary membrane and leads to failure in normal functioning, tissue ischaemia and organ damage. Interestingly, hyperglycaemia only displays an indirect association to vascular damage and there appears a group of patients that fail to develop microvasular complications even after long duration DM. This suggests a genetic basis for microvasulature damage (Shore and Tooke, '94).

1.6.4 DIABETES AND THE KIDNEY

End stage renal failure (ESRF) secondary to diabetic nephropathy is associated with significant mortality and morbidity. Diabetic nephropathy is now one of the most common causes of ESRF in developed countries and has considerable clinical and economic impact on dialysis and transplant programmes (Barnes et al., '98). Nephropathy develops in 35-45% of patients with IDDM and less than 20% with NIDDM. The process of nephropathy begins with the development of microabuminuria (urinary excretion of between 30-300mg/24hr) which some 5-10 years later progresses to overt proteinuria with urinary excretion of over 300mg/24hr. Hypertension almost

always develops during overt proteinuria and the decline into end stage renal disease begins, culminating in glomerular occlusion and complete loss of renal function (Andersen et al., '83, Krolewski et al., '85).

A characteristic set of structural abnormalities develops in the diabetic kidney including hypertrophy, an increase in the thickness of the basement membrane, an accumulation of mesangial matrix in the glomerulus, tubular atrophy and interstitial fibrosis. Variables related to IDDM induced kidney damage, as assessed by HbA1c assay, are primarily glucose exposure and to a lesser degree the duration of diabetes, but the greatest indicator of the probable development and progression of DN is hypertension (Parving et al., '81, Hasslacher et al., '85). Despite intensive glycaemic control there is currently no therapy that can reverse the progression of DN to ESRF (Barnes et al., '98). The risk of severe cardiovascular disease is 30-40 times higher for these patients than for diabetic patients without DN (Trevison et al., '97).

Because diabetic complications can be delayed and progression slowed by strict metabolic control, new forms of therapy are required for the effective treatment of DM. While control of blood glucose and anti-hypertensive treatment are current and effective treatments for DM complications, no standard treatment is effective at reversing the progression of DM pathology. Reversal of DN lesions has been achieved with whole pancreas transplants but this therapy is unrealistic as a comprehensive therapeutic solution when compared to the current demand (Barnes et al., '98, Fioretto et al., '98a, Fioretto et al., '98b). Continued investigations into the effects of hyperglycaemia and hypertension, in the setting of DM, will provide a greater understanding of the processes and underlying mechanisms. This in turn will lead to the identification of risk factors for DM complications and possible targets for new therapy.

1.7 MECHANISMS OF HYPERGLYCAEMIC STRESS

Hyperglycaemia specific disease in the retina, peripheral nerve, vasculature and glomerulus all share common pathophysiological characteristics (Brownlee, '92, Brownlee, '01). The order in which these changes occur is contentious but for simplicity they can be classified as:

1. Abnormal accumulation of glycated serum and cellular proteins, which persist and alter the normal function of either the native protein form or the ligand to which it can bind
2. The gradual expansion of extracellular matrix and increased permeability of extracellular and basement membranes.
3. A hypertrophic and/or hyperplasic effect on endothelial, mesangial and smooth muscle cells.

Each process involves metabolic or biochemical disturbances or altered gene transcription but appears more likely a combination of all three. While these are clear manifestations of glucose-induced pathology, there are likely to be various contributing factors as exposure to the same degree and duration of hyperglycaemia, manifest different susceptibility to damage (Krolewski et al., '88). What has emerged is a complex network of interrelated mechanisms and susceptibilities that may act in an autocrine or paracrine fashion, but which all contribute to the pathology of diabetes.

Currently there are five mechanisms by which glucose is believed to cause the processes described above and through them the pathophysiological changes seen in DM; these are increased polyol pathway flux, increases in the production of advanced glycation of serum and matrix proteins, activation/translocation of Protein kinase C isoforms, increased hexosamine pathway flux and an increase in oxidative stress. Addressing individual mechanisms, using targeted inhibition *in vivo,* has ameliorated one or some of the pathological complications, but there appears to be inter-related pathways that contribute to more than one mechanism and complete inhibition of all mechanisms remains elusive.

1.7.1 POLYOL PATHWAY

The polyol pathway follows the oxidation of intracellular D-glucose to sorbitol and then to fructose. The rate-limiting step in this pathway is the metabolism of glucose to sorbitol catalysed by the enzyme aldose reductase (AR), which is dependent on NADPH (illustrated in FIGURE 1.3). The second stage, which reduces sorbitol to fructose, is catalysed by sorbitol dehydrogenase (SD), and requires NAD^+. Aldose reductase has a low affinity for glucose and at physiological concentrations accounts for a very small amount of glucose consumption (reviewed (Nishikawa et al., '00, Lee et al., '00). The normal activity of aldose reductase is the detoxification of aldehydes to inactive alcohols. With the increase in intracellular glucose it is suggested that the pathway equilibrium is shifted to the production of larger amounts of sorbitol and fructose together with decreases in NADPH co-enzyme and increases in NADH co-enzyme.

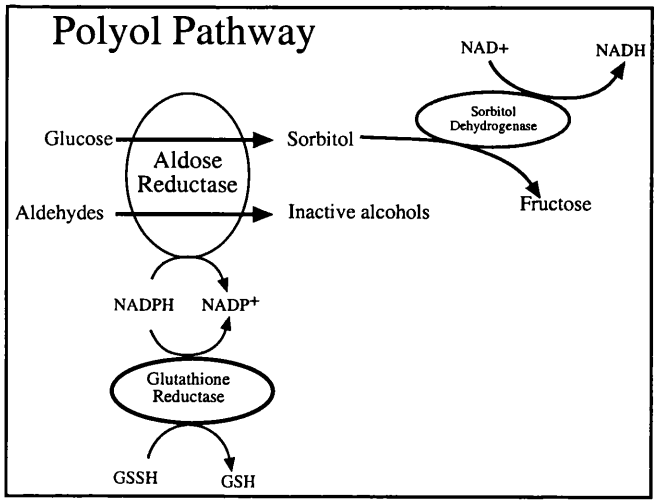


FIGURE 1.3. THE POLYOL PATHWAY.

Utilised as a removal mechanism for aldehydes, which includes glucose. The rate-limiting enzyme Aldose reductase (AR) begins the oxidation of cellular glucose to sorbitol, where it is then reduced to fructose by Sorbitol dehydrogenase. AR requires the cofactor NADPH, which links the polyol pathway reductive stress and the detoxification pathways requiring glutathione reductase.

The mechanisms by which the polyol pathway flux may contribute to hyperglycaemic damage may be due to intracellular hyper-osmotic stress and a shift in the redox balance of the cell, with NADPH depletion and increase in the cytosolic $NADH/NAD^+$ ratio, so called reductive stress. The hyper-osmotic theory is possibly tissue specific with cells of the retina and nervous tissue being more susceptible to osmotic stress than of the *vasa* or glomerulus (Feldman et al., '97, Hotta, '97). Experiments studying polyol formation, particularly sorbitol, have found that osmotic stress is only attributed to AR activation and sorbitol accumulation when cells are exposed to both hypertonic and hyperglycaemic conditions. (Bron et al., '93, Mizisin et al., '96, Mizisin et al., '97). Interestingly, experiments studying the reactive

intermediate Fructose-3-Phosphate (F3P) and its breakdown product 3-Deoxyglucosone (3-DG) in diabetic patients described a link between the polyol pathway and AGE formation (discussed later). Both F3P and 3DG can glycate proteins with very high efficiency and in these experiments the use of an aldose reductase inhibitor (ARI) reduced the level of AGE formation (Hamada et al., '96).

An increased NADH/NAD^+ ratio will inhibit GAPDH activity and favour increased concentrations of methylglyoxal (another AGE precursor) and diacyl glycerol (an activator of PKC) (Williamson et al., '93). NADPH is required for the regeneration of GSH from GSSH and thus cytosolic depletion could result in the generation or exacerbation of oxidative stress or failure of detoxification of reactive carbonyls. Reduced levels of GSH have been described in diabetic tissues and believed to be a result of flux through the polyol pathway (Morocutti et al., '98, Brownlee, '01).

Inhibition of aldose reductase *in vivo* has yielded inconsistent results. ARIs have been useful in inhibiting diabetic neuropathy but they have been of modest therapeutic value in the treatment of retinopathy, vasculopathy or nephropathy, possibly due to the loss of the primary detoxification function of AR (Engerman et al., '94, Greene et al., '99, Oates and Mylari, '99).

1.7.2 HEXOSAMINE PATHWAY

The hexosamine pathway diverts fructose-6-phosphate from glycolysis where it is converted into glucosamine-6-phosphate (GlucN-6-P), a reaction using glutamine as an amino donor and is catalysed by gluamine: fructose-6-phosphate aminotransferase (GFAT) as illustrated in FIGURE 1.4. Normally the hexosamine pathway accounts for between 2-5% of cellular glucose usage with the majority of glucose proceeding through glycolysis (McClain et al., '92, McClain and Crook, '96, Schleicher and Weigert, '00). GluN-6-P is rapidly converted further and then activated to UDP-GlucNAc for the provision of substrates in the formation of O-linked glycoproteins in the synthesis of cartilage and connective tissue. It is also suggested that protein glycolysis is analogous to phosphorylation in its importance in the activity of many transcription factors and cytosolic proteins. This links hexosamine flux and a number of cellular processes (Kreppel et al., '97, Hart, '97). The hexosamine pathway has also been implicated in insulin resistance with GFAT inhibition resulting in a 70% reduction

in insulin-responsive glucose transport GLUT4, (Marshall et al., '91b, Marshall et al., '91a, Chin et al., '97).

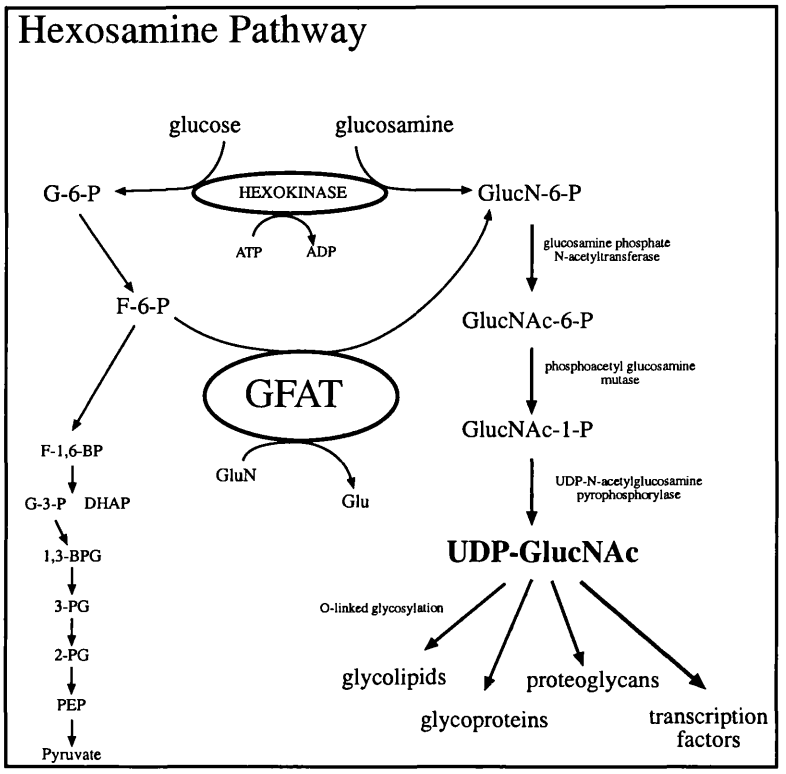


FIGURE 1.4 THE HEXOSAMINE PATHWAY.

The removal of the glycolysis intermediate fructose-6-phosphate (F-3-P) into the hexosamine pathway involves the rate-limiting enzyme GFAT (see text). This pathway is utilised for the provisions of glycation substrates for a variety of cellular proteins including, but not limited to cartilage genesis, lipids and transcription factors. The flux of glucose through glycolysis is believed to shunt this pathway in the forward direction to over production of glycation substrates used in several cellular processes.

The inhibition of GFAT *in vitro* blocks the high glucose induced increase in transcription of TGF-α, TGF-β1 and PAI-1 (Kolm-Litty et al., '98a). The precise mechanism on how flux through the hexosamine pathway effects changes in gene transcription is unclear but several explanations have been proposed. Treatment of cultured cells with either glucose, glucosamine or indeed over expression of recombinant GFAT constructs produces increases in gene transcription and alters regulation of PKC isoforms (Kolm-Litty et al., '98b, James et al., '00, Goldberg et al., '00).

The glycosylated form of Sp1 appears to be more active in transcription that the un-glycosylated form (Kadonaga et al., '88), and glucosamine was shown to activate the PAI-1 promoter through Sp1 in glomerular mesangial cells (Chen et al., '98b, Goldberg et al., '00, James et al., '00). Nearly all RNA polymerase II transcription factors are O-acetylglucoasamine modified (Hart, '97), and thus it is reasonable to expect that transcription factors other than Sp1 are modified through the hexosamine flux. This

pathway may therefore represent a less specific and more generalised mechanism of transcription altered by hyperglycaemia (Du et al., '00).

Transcription factors are only one family of proteins that are modified by O-acetylglucosamine; cytoplasmic and other nuclear proteins can also be modified and or their activity controlled in this manner (Kreppel et al., '97, Hart, '97). eNOS and PKC isoforms are also susceptible to O-acetylglucosamylation and, once modified, alter behaviour such as translocation and substrate specificity (Kolm-Litty et al., '98b, Du et al., '01). Activation state, like the dissociation of latent TGFβ1 to its active form is also affected by treatment of cells with glucosamine (Kolm-Litty et al., '98a). Over-expression of GFAT has also been shown in oxidative stress although the precise mechanisms remain elusive (Kaneto et al., '01).

From this evidence it seems likely that flux through the hexosamine pathway may represent a mechanism by which hyperglycaemia may mediate altered gene transcription as well as metabolic disturbances.

1.7.3 AGE FORMATION AND PERSISTENCE

Advanced glycation end products (AGEs) are the result of non-enzymatic reactions between glucose (and its metabolites) and protein amino groups (see FIGURE 1.5). Accumulation of AGE in tissue is associated with age and chronic illness such as diabetes (Brownlee, '95). Originally identified and traced by the ability of certain AGEs to 'brown' skin tissue or their fluorescence under spectral excitation, specific and more accurate AGE markers continue to emerge and positive association with diabetes remains strong. AGEs are found in increased amounts in the diabetic retina, vasculature, skin and glomeruli but their generation and ubiquitous nature suggests they will be present in other tissues (see FIGURE 1.5).

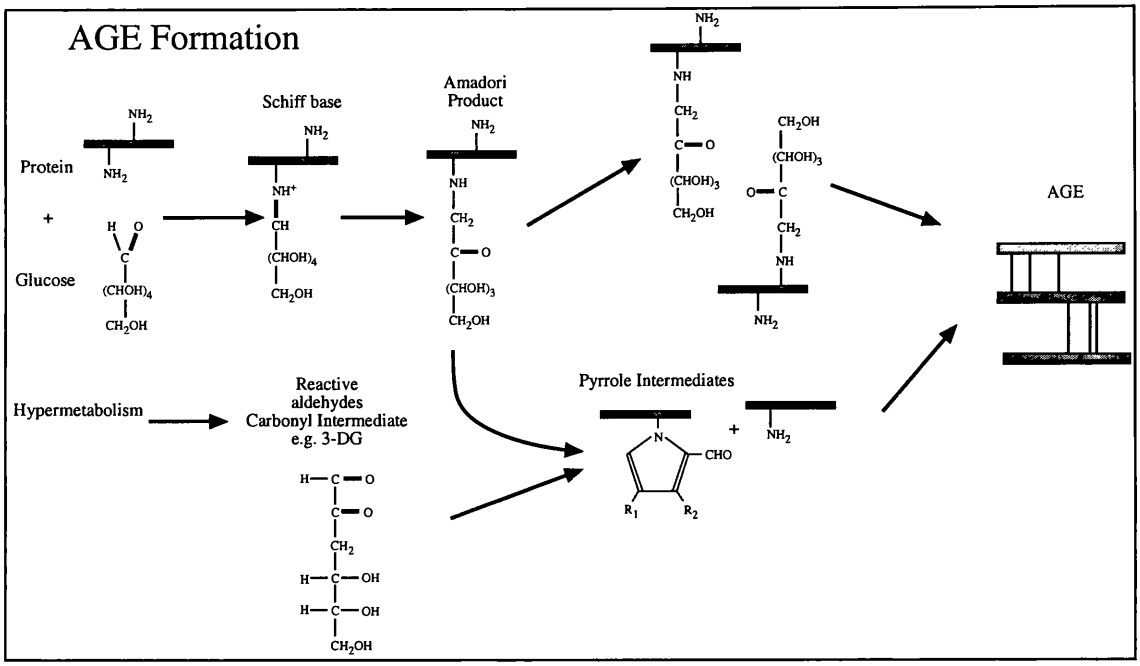


FIGURE 1.5. PRODUCTION OF AGES FROM GLUCOSE.

Production of AGEs can progress through two possible routes. First in the Mailliard reactions where extra-cellular glucose reacts, non-enzymatically, with protein motifs and create Amadori and pyrrole products via intermediate Schiff bases. Re-arrangement of these products result in cross-linking of proteins and persistence due to retarded turnover. Secondly, and believed to be more important in DM, is the intracellular generation of reactive carbonyl intermediates in a hyper-metabolic state. These intermediates will also cross-link proteins and retard their turnover. Both processes are non-specific for target proteins and can alter or destroy a proteins function.

AGEs are damaged proteins. The diverse nature of AGEs in the diabetic setting means that any pathological significance is complex. Currently AGEs contribute to pathology through three mechanisms. Firstly, altered proteins, either intracellular or extracellular, may possess altered function and as glycation is only specific to amino groups, these altered functions may not be specific. The second mechanism is the glycation of matrix proteins. Matrix proteins by their nature are longer lived and any detrimental effect to matrix-matrix or matrix-cell interaction will persist. Finally, AGEs possess specific binding proteins, presumably as a detoxification process for normal turnover. Receptor mediated AGE signalling has been described and associated with the cellular events that have previously been attributed to glucose namely ECM expansion and albuminuria in DN (Pugliese et al., '97, Ziyadeh et al., '98). These actions are blocked with inhibitors of AGE formation. Additionally there have been reports of polymorphisms within the gene for RAGE and thus a possible genetic susceptibility to AGE dependant pathology (Hudson et al., '98). Much research has been directed at delineating the action of glucose *per se* and the actions of glucose altered proteins.

1.7.4 PKC ACTIVATION

Protein kinase C incorporates a family of enzymes that play fundamental roles in many cellular signal transduction pathways and specific roles in cellular function. Signal transduction involving PKC isoforms can result from altered phosphorylation states as well as translocation to areas within the cell (Dempsey et al., '00). PKC isozymes are involved in a variety of cellular functions including permeability, cellular contraction, migration, proliferation, hypertrophy, apoptosis, secretion pathways and gene transcription. PKC isozymes are broadly classified into families based on their activation requirements and homology in specific regulatory regions (see TABLE 1.2).

Family subgroup	Isoform	Tissue distribution	Co-Factor requirement		
			Ca^{++}	DAG	PS
Conventional cPKC	α	Widespread	+	+	+
	β_I	Widespread	+	+	+
	β_{II}	Widespread	+	+	+
	γ	Brain	+	+	+
Novel nPKC	δ	Widespread	-	+	+
	ε	Brain, heart	-	+	+
	η	Heart, lung, skin	-	+	+
	θ	Muscle, brain blood cells	-	+	+
	μ	Lung epithelial cells	-	+	+
Atypical aPKC	ζ	Widespread	-	-	+
	ι/λ	Kidney, brain, pancreas	-	-	+

TABLE 1.2. PROTEIN KINASE C ISOFORMS.

Classical PKC isozymes (cPKC) are dependent on Ca^{++} and phospholipids. A second group, novel PKC (nPKC), do not require Ca^{++} but remain dependent on phospholipids. The final group classified thus far are the atypical PKCs (aPKC); members of this group do not require phospholipid or Ca^{++} but appear to require phosphatidyl serine (PS). An activator of classical PKC isozymes is the diglyceride diacylglyceride (DAG) that is primarily created by the cleavage of phosphatidylcholine by phospholipases (see FIGURE 1.4). DAG is also formed through the intracellular signalling pathway that cleaves phosphoinosityl lipid PIP_2 into DAG and IP_3 (Schnaper, '00, Black, '00).

Intracellular hyperglycaemia increases the level of DAG in cultured cells as well as the retina and glomeruli of diabetic rats. Glucose may be directly metabolised to DAG through triose phosphates created in glycolysis and their utilisation in the pentose phosphate pathway (PPP) (see FIGURE 1.6) (Wolf et al., '91, Xia et al., '94). This *de novo* synthesis from the glycolytic intermediates, as opposed to PIP_2 cleavage, may also be exacerbated by the high glucose induced increase in $NADP^+$ from the polyol pathway, which will drive the PPP. Increased DAG activates at least 9 of the 11 PKC isoforms. Most prominent is the activation of beta (β) and delta (δ) isoforms but

epsilon (ε), zeta (ζ) and alpha (α) isoforms are also activated in what appears a tissue/cell specific manner (Xia et al., '94, Koya et al., '97).

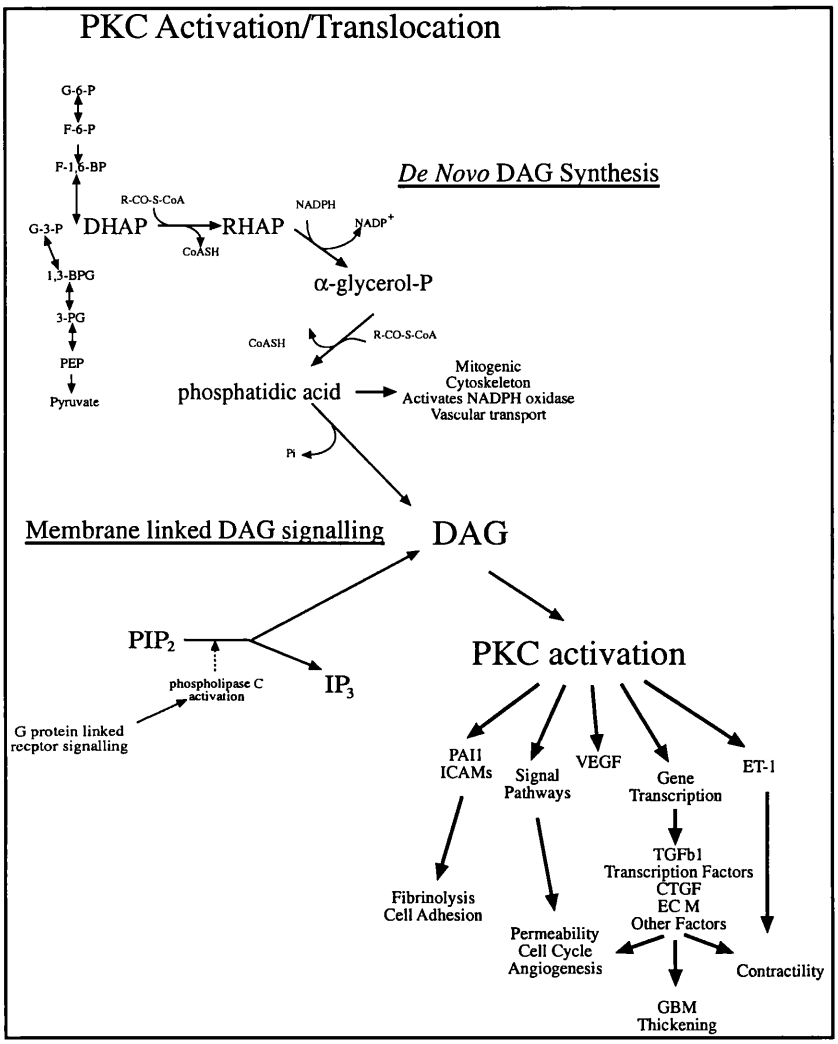


FIGURE 1.6.
ACTIVATION AND ACTIONS OF PKC.
DAG can activate various isoforms of PKC. The production of DAG can increase through the splicing of DAG from PIP_2 during membrane-linked secondary messaging systems as well as the *de novo* production from glycolytic intermediates. PKC can then take part in many cellular processes through the activation of active sites or the translocation to specific cellular compartments.

In diabetic models, PKC activation and translocation has been implicated in abnormal blood flow and vascular permeability, generally mediated through the actions of ET-1 (PKC-β), VEGF (PKC-α) and PDGF (Ishii et al., '96, Williams et al., '97, Hempel et al., '97). Culturing glomerular mesangial cells in high glucose altered the translocation dynamics of PKC-δ and ε from the cytosol–to-membrane cellular fraction to the cytosol-to-particulate / nuclear or cytoskeleton cellular fraction. This translocation changed the way that the cultured cells responded to the vaso-active peptides ET-1 and PDGF-b (Glogowski et al., '99). MC and vascular smooth muscle cells (VSMC) share similar lineages and are involved in glomerular and vascular haemodynamics.

1.7.5 OXIDATIVE STRESS

Oxidative stress is defined as a tissue injury produced as the result of increases in reactive oxygen species (ROS) such as the superoxide anion ($O_2^{\bullet-}$) and hydroxyl radical ($^{\bullet}OH$). Evidence to support a role for oxidative stress in IDDM is found in the use of antioxidants to suppress the high glucose induced changes seen in cultured mesangial and endothelial cells (Trachtman et al., '93, Trachtman, '94, Ha et al., '97). It is unclear whether oxidative stress is an important link between hyperglycaemia and cellular disturbances or whether it is a secondary consequence of the primary pathogenic mechanisms. Certainly ROS have a functional role in all the above mechanisms and addressing the oxidative stress of a cell in hyperglycaemic stress ameliorates many of the metabolic disturbances (Giugliano et al., '96, Baynes and Thorpe, '99, Ha and Lee, '00, Brownlee, '01).

It has been proposed that oxidative stress in a hyperglycaemic state is mediated by the production of superoxide from metabolic oxidative phosphorylation on the electron transport chains and that this is due to the increased reductive stress created by flux of glucose through glycolysis. Reducing equivalents from glycolysis and the TCA cycle, are transferred to the transport chain by way of NADH and $FADH_2$ to complexes I and II respectively. When the proton gradient is high, complex III in the electron transport chain is inhibited and the life of ROS producing intermediates, such as ubisemiquinone in the 'Q' cycle, is prolonged (see FIGURE 1.7). Oxidative stress can then affect various cellular functions but one, inhibition of GAPDH, is proposed to be important in the manifestation of hyperglycaemic disturbance. Inhibition of GAPDH can reach 66% in hyperglycaemia and the depletion of glycolysis intermediates into each of the four mechanistic pathways, polyol pathway flux, hexosamine flux, *de novo* DAG dependent PKC activation and AGE formation are all upstream of GAPDH (Du et al., '00).

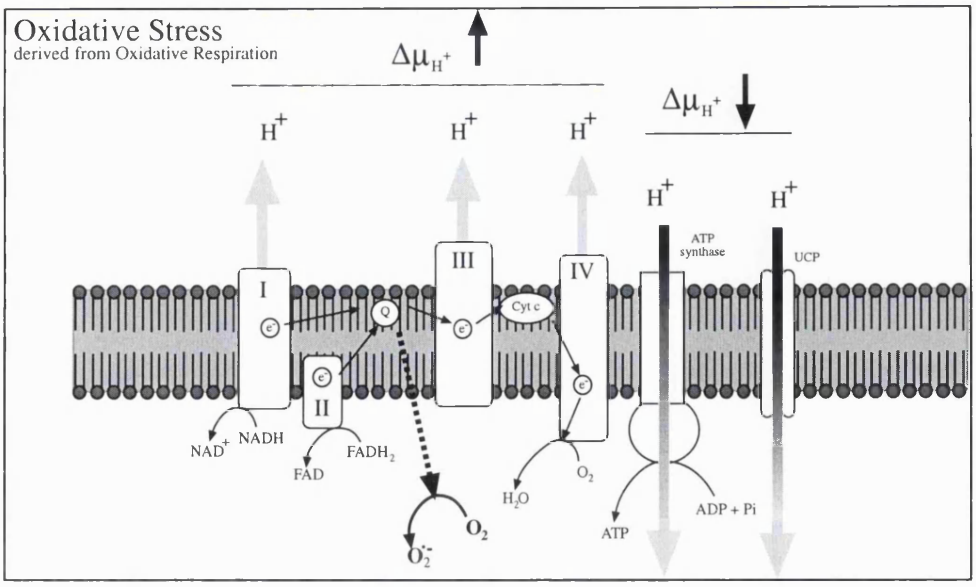


FIGURE 1.7. ELECTRON TRANSPORT IN OXIDATIVE PHOSPHORYLATION.

Reducing equivalents ($NADH/FADH_2$) are transferred to subunits I & II in electron transport. Passage to lower free energy results in the translocation of electrons $\Delta\mu_{H^+}$. ATPase utilises this potential difference to phosphorlyate ADP to ATP and the whole gradient can be collapsed with the uncoupling protein 'UCP'. The transport is inhibited when $\Delta\mu_{H^+}$ rises above a threshold level and this blocks transport at complex III. In a hypermetabolic state this results in the persistence of reactive intermediates, which can pass reducing power onto molecular oxygen forming reactive superoxide. $O_2^{\bullet}$. Superoxide will increase the oxidative burden of the cell and inhibit GAPDH in glycolysis. GAPDH inhibition will shunt glycolytic intermediates into all the pathways associated with DM.

Contrary to the concept of oxidative stress as a primary mechanism is the hypothesis that aspects of carbonyl stress are independent of oxidative stress yet still contributory to diabetic pathology (Baynes, '91, Baynes and Thorpe, '99, Suzuki and Miyata, '99). The distinction between anti-oxidation and detoxification must be considered when addressing this situation. GSH has both antioxidant and detoxification activities. Methylglyoxal (MGO) is increased in diabetic patients and contributes to the formation of AGEs and is detoxified through GSH dependant mechanisms. MGO is formed by non-oxidative means primarily from triose phosphate intermediates in glycolysis. The generation and subsequent disposal of MGO is therefore independent of ROS, although they may be intermediates, and as such cannot be a consequence of oxidative stress, but rather is a consequence of overload or failure of the detoxification pathway.

1.7.6 THE MECHANISMS FOR HYPERGLYCAEMIA INDUCED PATHOLOGY REMAIN COMPLEX

From the above review of glucose induced disturbances it is apparent that the numerous mechanisms form part of a complex network of pathways. The net result is the pathology associated with DM. The differences through which hyperglycaemia effects cells of various phenotypes, and more importantly targeted pathways in the various cells, is equally complex. Drugs that target the oxidative stress of endothelial cells of the *vasa* appear not so useful in the glomerulus. AR inhibitors (ARIs) showed modest therapeutic value for treatment of neuropathy but are of little use in the nephropathy. Clearly, a unifying mechanism preceding all other mechanisms is an attractive hypothesis. Currently such a mechanism is absent and thus a single effective therapy is absent. The glucose hypothesis persists, but intensive insulin treatment as a therapy to address hyperglycaemia can only slow rather than reverse the progression of DM pathology. As such, effective treatments for particular complications are still required.

1.8 GLOMERULAR APPARATUS IN DM

1.8.1 THE GLOMERULUS

The glomerulus is the primary filtration unit of the kidney (see FIGURE 1.8). Blood is supplied to the glomerular apparatus through the afferent arteriole and the plasma is filtered through fenestrations in the endothelial cells and the glomerular basement membrane, then through the pores between the podocyte foot processes (slit pores) into Bowman's space. The capillary endothelium, GBM and slit pores constitute the filtration barrier that collectively filter the plasma allowing passage of water, electrolytes, metabolic waste, small proteins and small organic molecules through to the nephron. The filtered blood is drained from the glomerular apparatus through the efferent arteriole. The glomerular capillary wall consists of three layers: an endothelial layer, a basement membrane (GBM) and an external layer of epithelial cells. The GBM is continuous throughout the glomerulus, surrounding each capillary loop and 'stalk' region of the glomerular tuft. At the vascular pole the GBM is continuous with Bowman's capsule. Filtration through the GBM is of primary importance to the

function of the glomerulus and perturbations in the filtration rates or porosity are associated with many renal diseases (Kriz and Kaissling, '92, Tisher and Madsen, '00).

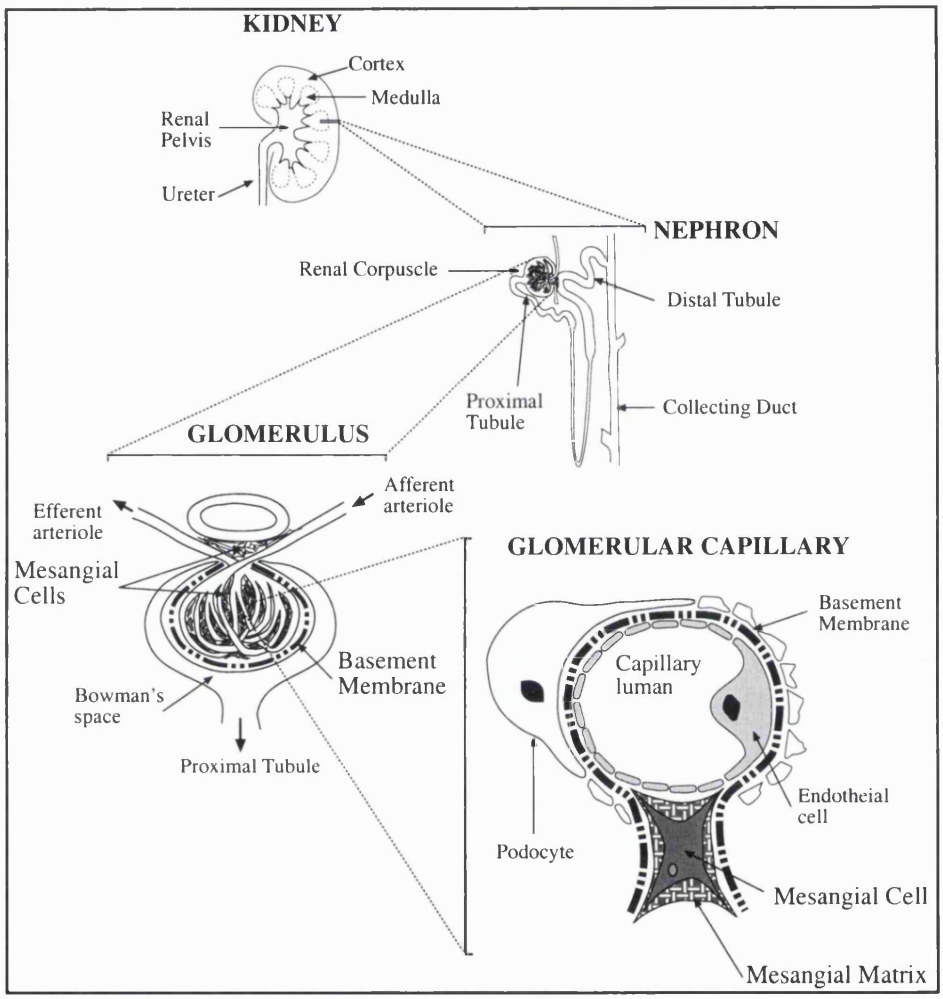


FIGURE 1.8. THE GLOMERULUS.

Occupying the cortex and medulla of the kidney are the nephrons. The renal corpuscle forms the major filtration unit of the kidney. Blood is supplied through the afferent arteriole, is filtered in the glomerulus and exits through the efferent arteriole. Filtrate passes through the capillary fenestration and glomerular basement membrane, collecting in Bowman's space before passage into the proximal tubule. The 'stalk' of the glomerulus is composed glomerular mesangial cells, which play structural and functional roles in glomerular function. The mesangial matrix surrounds the mesangial cells and uncontrolled expansion is a histological marker of much renal pathology.

1.8.2 HISTOLOGY OF THE GLOMERULUS

Three types of cell are found within the glomerular tuft. Endothelial cells line the capillary luminae inside the GBM and mesangial cells reside within the stalk region surrounded by mesangial matrix (mECM) (see FIGURE 1.6). The external surface of the GBM is covered in podocytes or visceral endothelial cells that are continuous to the vascular pole with the epithelial cells of Bowman's capsule. Filtration through the

glomerulus requires movement through the highly fenestrated capillary endothelial layer, the porous junctions of the podocytes and the GBM. The capillary endothelial layer does not impede passage of large molecules and so the GBM and porous junction of the podocytes provide the stringent filtering of the glomerular apparatus. Movement through the GBM and podocyte junctions is related to the size shape and charge of any macromolecule.

The central area of the glomerulus is predominantly composed of the mesangium and is surrounded by endothelial cells and capillary networks. Much of the original work describing the mesangium concerned the ultra-structural appearance of cross-sections of glomerulus using electron and light microscopy. Current understanding is that the mesangium recognises a multifunctional entity and mesangila cells are capable of various tasks. These include, but not limited to, structural support, construction and turnover of mesangial matrix, producing and target cells for vasoactive agents such as angiotensin II (AngII), prostaglandin E2 (PGE2) or nitric oxide (NO), for inflammatory mediators, cytokines and mitogens, and the biological handling and clearance of macromolecules like lipid, immune complexes and AGEs (Risdon, '85, Trevison et al., '97).

1.9 GLOMERULAR MESANGIAL CELLS

1.9.1 FUNCTIONS OF MESANGIAL CELLS

Similar to vascular smooth muscle cells (VSMC) the mesangial cell contains large amounts of actins, myosins and tropomyosins indicative of a contractile function. Receptors for vasoactive peptide angiotensin II (AngII) support this role (reviewed in (Kreisberg et al., '85)). This suggests a role for MCs in the regulation of glomerular haemodynamics. ECM surrounds, is produced and remodelled by the MC, so they also function to create a structural framework for the glomerular tuft. Additionally the MC has been reported to possess phagocytic properties that implicate them in the uptake and clearance of macromolecules from the glomerulus (Schreiner et al., '81). Inappropriate activation of any of these functions, particularly the contractile and ECM remodelling, are common observations in many renal diseases. An increase in local cellular population, independent of cellular infiltration, has been shown to be a result of mesangial cell proliferation. Expansion of the mesangial ECM, both due to the

increased production of matrix proteins together with decreased degradation, is also seen in renal disease.

Estimates of proliferation of MC *in vivo* describe a low rate of about 1% per day (Pabst and Sterzel, '83). The control of MC proliferation appears to play an early role in the progression of glomerular pathology. Matrix expansion is generally preceded by MC proliferation in models of glomerular nephritis. Persistent MC hyperplasia caused by repeated injury is believed to lead to irreversible scarring and the eventual loss of glomerular function (Couser, '93, Floege et al., '93). Unlike normal wound healing, the proliferative responses of MC are presumed to be a necessary physiological response for the re-constitution of renal tissue. Models of anti-Thy1.1 nephritis exhibit marked proliferation of MCs and expansion of the mECM. With no successive insult, resolution of the MC hyperplasia and return of glomerular function is observed. A similar phenomenon is seen in Habu toxin glomerulopathy or inflammatory disease such as post-streptococcal glomerulornephritis. Understanding the roles of MC proliferation and mECM appear more complex than first thought. Reducing MC proliferation is associated with reduced scarring but if attenuating the proliferative response of MC to injury in some way results in defective repair and insufficient cellular re-population, then clearly a more thorough strategy is required. Discriminating between reconstructive proliferation of MC and non-reconstructive activation of mesangial cells seems essential to an evaluation of possible intervention therapy. Theoretically one can rationalise the function of a MC, indeed any cell, by its vulnerability and response to stimuli. MCs are known to possess receptors for a variety of growth factors, mitogens and other biologically active peptides. Study of the manner in which MC respond to these factors will assist in discerning the nature of disease model and thus indicate a more suitable intervention. Currently there is no reliable method to distinguish between these two processes *in vivo* and so a complete understanding of the native mesangial cell remains elusive. With the advent of *in vitro* culturing of MC a large amount of data regarding specific cellular responses to many stimuli is being generated (Mene et al., '89, Davies, '94).

1.9.2 HISTOLOGY OF MESANGIAL CELLS

The contractile ability of the glomerular mesangial cell under certain conditions has lead to the MC being considered a type of vascular contractile pericyte (Mene et al., '01). The stellate morphology of the MC comprises long processes that reach into the matrix or neighbouring cells. The MCs are arranged circumferentially to form loops around capillaries where they can control the GFR through contraction and relaxation. The MC is rich in contractile cytoskeletal elements and the distribution of actinomyosin is similar to the smooth muscle cells of the vasculature. Actin bundles span the entire length of the mesangial cell processes and establish contact with various anchoring points on neighbouring cells and the ECM that surround the cell. Through this cellular network, the mesangium connects the glomerular basement membrane with the endothelial capillary lining, thus providing a link controlling the glomerular filtration. The mesangial cell establishes connections to the mECM through fibronectin, laminins and integrins. The mECM is composed of a dense network of elastic micro fibrils similar to the connective tissue of many organs, yet also contains a network of intercellular channels that can traffic macromolecules. MC's produce a variety of matrix proteins and thus are also believed to be involved in the construction and remodelling of the mesangial ECM. Thus, it would appear that the multifunctional mesangial cell plays a central role in the correct functioning of the glomerulus, in regard to both the mechanical integrity and the dynamics of glomerular filtration. The MC also appears to possess phagocytic activity and is also capable of the production and secretion of a variety of growth factors that can act locally or elsewhere producing various responses (reviewed in (Davies, '94).

For the glomerulus to function correctly it is necessary for these functions to be tightly regulated and it is the augmentation of these activities, either in concert or individually, which characterise many renal diseases. Because of these structural characteristics, mesangial cells have been referred to as myofibroblasts, cells that have features of smooth muscle cells and fibroblasts. Additionally because of the MC ability to alter phenotype in order to perform specific structural metabolic and secretory functions, it may be thought of as a pluripotent interstitial cell of the glomerulus, which changes function when stimulated in a specific manner. The presence of receptors for a variety of vasoactive, mitogenic factors, chemokines and cytokines support this hypothesis (reviewed in (Mene et al., '89).

Mesangial cells contain many of the same contractile proteins as smooth muscle cells but some isoforms of these contractile proteins are specific for SMC or MC (reviewed in (Stockand and Sansom, '98). Despite confusion regarding specific isoforms present in each cell type, consensus remains that both possess functioning contractile apparatus. Contraction in MCs is initiated by Ca^{++} release and the activation of myosin light chain kinase in a calmodulin-dependant fashion. In contrast to skeletal muscle, and similar to SMC, the contractile filaments in MC traverse the stellate processes from one side to the other and are not aligned in longitudinal bundles. Actin containing microfilaments connect the endothelium and GBM to the cell and are the predominant microfilaments in MC's. Bundles of extra cellular micro fibrils form an extensive interwoven 3D meshwork that provides many anchor points between the MC mECM and GBM. This structure allows for the simultaneous support of the mECM and glomerular capillaries, together with control of the capillary surface area. Control of GFR has been described as both static and dynamic. Static control involves the MC detecting increased capillary pressure and countering this distension of the GBM with a contraction. In this way the GFR is maintained, countering the expansion of the mesangium with a reduction in capillary surface area. Dynamic control of the GFR is also evident in experiments using vasoactive peptides. Mesangial cells posses receptors and ligands for vasoactive substances that both initiate and antagonise contraction. Factors such as angiotensin II (Ang-II), endothelin-I (ET-I) and arginine vasopressin (AVP) initiate strong contraction signals in cultured glomerular MC, which mimic the *in vivo* contraction signal. In addition to these factors, growth factors like PDGF and platelet activating factor (PAF) will produce a mitogenic response as well as a cellular contraction. On the contrary, atrial natriuretic peptide (ANP) and nitric oxide (NO) produce powerful relaxation signals in cultured MC and have been shown *in vivo* to cause increases in GFR.

Much information regarding the function of the mesangial cell and responses to a variety of stimuli has been described *in vitro*. Availability of stable culture of primary mesangial cells and immortalised mesangial cell lines (HMCL) has accelerated understanding of the MC (Sraer et al., '96). Currently, the mesangial cell has become a respected model for studying various aspects of glomerular disease. Demonstrations of cellular activities associated with pathology have been described using mesangial cells such as the effect of shear and stretch forces on the activity and production of

vasoactive peptides such as AngII and NO (Nagata et al., '92, Cortes et al., '97). Stimulating mesangial cells by culturing them in the presence of activators of intracellular signal cascades, like those associated with RAGE and PKC activation, have demonstrated the generation of specific alterations in metabolic and transcription activities (Yan et al., '94, Schmidt et al., '95). Culturing the mesangial cell under high concentrations of glucose has been fundamental in the elucidation of the many pathways described above and resulted in the formulation of the current paradigm of the effect of glucose on the mesangial cell in the context of diabetic nephropathy.

1.9.3 MC AND THE MESANGIAL MATRIX (MECM)

The mesangial matrix represents a mass of information regarding the regulation of cell behaviour. Alteration in the quantity and composition of mECM has profound effects on MC biology and may alter the proliferative response of MC (Floege et al., '90, Marx et al., '93). The mECM exists as a balance of synthesis and degradation. Net deposition of matrix proteins results from disturbances in this balance and is a common feature in glomerular sclerosis. MCs seeded on collagen or fibronectin coated culture flasks display a greater proliferative rate than those seeded on plastic alone. Treating cultures with thrombospondin and heparin sulphate has demonstrated inhibition of MC proliferation *in vitro*. In addition, many matrix proteins are able to sequester mitogenic factors like TGFβ and bFGF and create a local store of inactive but available factors.

The attachment of cells to ECM is an important factor in the cell cycle. Activation of cyclins D1 and E/cdk2 complex is anchorage dependent in fibroblasts (Iyer et al., '99). Integrins and their ligands also take part in cell cycle progression and constitute a major component to the mECM. The exact mechanism of action for integrins appears complex as the binding to some integrins induces a proliferative response while the binding to other integrins produce an inhibition of proliferation. The production of collagen type I, normally not noted in glomeruli, is present in certain forms of glomerulonephritis (Floege et al., '91, Yagame et al., '95). Growing MC on different substratum of collagen type I produced striking differences in proliferative state. Collagen I monomers did not affect the proliferation of MC, but polymerised type I collagen completely prevented MC proliferation (Schocklmann et al., '99, Schocklmann et al., '00). In summary the growth characteristics and cellular responses depend upon the substrata to which the cells bind.

1.10 MECHANISMS OF MESANGIAL CELL DYSFUNCTION

Some of the pathologic mechanisms present in DM are present in MC. Under the conditions that mimic DM, low insulin and high glucose, the glomerular mesangial cell undergoes a peculiar growth pattern (Steffes et al., '89, Osterby et al., '90). Initially the MC enters a proliferative stage, which is followed by a period of growth arrest where the cells appear locked at a particular point of the cell cycle. From this point the cells become hypertrophic and irreversible structural changes begin. The MC begins to secrete a variety of cellular factors and matrix proteins and initiate altered transcription of genes involved in a number of cellular processes including the cell cycle, matrix turnover and metabolism.

1.10.1 CELLULAR FACTORS

Much research demonstrated that several growth regulation proteins affect the tone of the glomerular mesangial cell as well as the deposition and composition of the mesangial matrix (Mene et al., '89, Striker et al., '91, Nakamura et al., '93). Mesangial cells respond to a variety of growth factors and other cellular peptide stimuli and alter their growth pattern and matrix modelling (see TABLE 1.3). The expression of these proteins as paracrine activators has revealed that they have the ability to modulate their own gene expression and that in neighbouring cells. The altered transcription of a marker of proliferation, PCNA, inflammatory cytokines, TNF-α, and PDGF-bb, a pro-fibrotic cytokine, TGFβ1, and basic fibroblast growth factor (bFGF) increased in the glomeruli of STZ-induced diabetes in rats. The modulation in transcription of the factors was only partially ameliorated by insulin treatment (Nakamura et al., '93). This provided evidence that DM causes the disruption of the proliferative state and induced a pro-fibrotic environment in the glomerulus, which appeared to be a result of the activation of various growth factors. These cellular factors were capable of increasing the expression of matrix molecules in mesangial cells as well as increasing the proliferation of mesangial cells.

Increases in the expression of TGFβ have been demonstrated in human and animal models of DN as well as cultured primary and transformed mesangial cells (Nakamura et al., '93). Stimulation of TGFβ1 expression has been demonstrated in all the classical mechanisms of DN from high ambient glucose and the downstream metabolic pathways, AGE stimulation, vasoactive peptides and mechanical strain (Cooper, '98). Moreover, treatment with neutralizing anti bodies directed at TGFβ has provided strong evidence for a causal link between TGFβ and structural and functional abnormalities of DN. (Ziyadeh et al., '00) described long term treatment of diabetic mice with anti-TGFβ antibody, which prevented renal insufficiency, excess matrix expression and glomerular matrix expansion, the main features of diabetic nephropathy. A causal relationship between the intra-renal TGFβ system and matrix expansion was demonstrated.

Growth factor	Modulation DN or High Glucose	Translation Transcription	Time Scale	Ref
PCNA TNF-α PDGF-a/ab TGFβ IGF-1 BFGF EGF	MRNA for PCNA, PDGF-b, TNF-α, TGF-β and bFGF were all increased in the glomeruli of STZ induced DM. mRNA for IGF-1, PDGF-a and EGF did not alter in the same model			(Nakamura et al., '93)
TGFβ1	Persistent increases across chronic exposure to glucose	Bio availability and mRNA	21 days	(Yamamoto et al., '93, Studer et al., '95)
ET-1	In high glucose ET-1 PKC translocation is altered	-	48hrs	(Glogowski et al., '99)
CTGF	Appears increased in chronic High Glucose	Transcription	7 days	(Murphy et al., '99)
VEGF	Increased mRNA but unclear is downstream action is paracrine or autocrine	Transcription	3hrs	Cha et al '00

TABLE 1.3. CELLULAR FACTORS IMPLICATED IN DN BY ASSOCIATION WITH MC.

Not all factors associated with MC induced DN involve transcription. However many of the initiating factors are associated with increased transcription in *in vivo* and *in vitro* model systems. The classic factor associated with DN is TGFβ1. Both directly contribute to the expansion of mesangial ECM, the histological hallmark of DN.

Other factors that have attracted attention have been ET-1, Renin-Angiotensin (RA) system, IGF and associated binding proteins (VEGF, CTGF and PDGF) but it is unclear whether the altered expression of the factors are downstream of the TGFβ1 system or inducers (see TABLE 1.3). Certainly, the independence of the haemodynamic

mechanisms from the TGFβ1 system suggests a separate role for vasoactive peptides at least.

1.10.2 EXTRACELLULAR MATRIX FACTORS

The expansion of the mesangium is the histological hallmark of DN and has been extensively studied. Much work has involved determining the primary causes in the increased expression of ECM molecules but it was also noticed that a reduced turnover, as well as increased secretion, could play a role in the net expansion of the mECM (Mclennan 94). Matrix turnover involves three factors, matrix proteases (MMPs), plasmins that activate latent MMPs and inhibitors of MMPs (TIMPS). High glucose and DM has been shown to modulate MMP expression in a complex fashion. MMP-2 appears increased in high glucose MC while MMP-9 and-7 appeared to decrease. The controlled expression of TIMPs appears equally complex (McLennan et al., '94, Shankland et al., '96, Fisher et al., '97).

Matrix Factor	Modulation in DN or HG	Level of detection	Time scale	Ref
Collagens	Type IV increased. Type I & III not normally expressed	mRNA and plating efficiency	Chronic >96hrs	(Simonson et al., '89, Yamamoto et al., '93)
Laminin	Increased 60%	Protein ELISA	4 weeks	(Ayo et al., '91a)
Integrins	Increased b1, a3/a5 no change	Immunodetection	24hrs	(Ihm et al., '95, Ayo et al., '90, Ayo et al., '91a)
Fibronectin	In culture mesangial cells secrete more Fibronectin (FN) into the supernatant and the matrix in HG.	ELISA	14 day	(Nahman et al., '92)
Thrombospondin	2.3 fold increase	mRNA	21 days	(Holmes et al., '97)
ICAM/VCAM	No change	mRNA	24hrs	(Ihm et al., '95)

TABLE 1.4. MODULATION OF MC MATRIX PROTEINS IN DN OR HIGH GLUCOSE.

Expansion of the mesangial ECM is the histological hallmark of DN and *in vitro* models have implicated many ECM components in DN. Matrix components, particularly collagens, laminins and integrins constitute a large group of matrix factors. The interaction between a cell and the matrix that surrounds it is complex. The matrix is able to alter phenotype and influence transcription through a series of cellular receptors. Taken together these reasons form the basis of an entire study of cellular environment that include MC grown in high glucose.

The changes in the ECM, both in net increases and composition, has a pseudo-auotcrine affect on MC. The ligands for matrix proteins, integrins and other adhesion molecules, have the ability to initiate and transduce signal themselves, presumably

through a mechanism that senses cell contact or stretch functions (Sterzel et al., '92, Rupprecht et al., '96, Brady et al., '00). It thus follows that any change in the composition of the ECM will alter the way in which the MC initiates and transduces such signals (Makino et al., '93, Makino et al., '95). Altered expression of integrins and other ECM molecules has been extensively described in MC grown in high glucose as well as *in vivo* models (see TABLEs 1.4 & 1.5)

Proteinase/Inhibitor	Modulation in DN or HG	Ref
Plasmins	Decreased in DM and MC under high glucose	(Fisher et al., '97)
MMPs	Generally increased mRNA in MC and DM but specific indication of repression may exist	(McLennan et al., '98)
TIMPs	Generally increased	(Murphy, '95, Shankland et al., '96)

TABLE 1.5. MATRIX TURNOVER FACTORS AND THE EFFECT OF DN OR HG.

The matrix exists as a balance between deposition and degradation. In periods of stress or altered phenotype matrix remodelling occurs and requires the removal of redundant matrix components. The importance of mesangial ECM in DN the effect of glucose on enzymes that degrade matrix proteins has been investigated. There appeared a complex pattern of suppression and activation among the degrading enzymes MMP and inhibitors of MMPs TIMPs. This indicates that the balance of matrix remodelling is disturbed during glycaemic stress.

1.10.3 THE CELL-CYCLE

The expansion of the ECM has been at least partially attributed to the disruption of the cell cycle in the HMCL (Wolf et al., '97). It was observed that culturing HMCL in the presence of high glucose resulted in a significant reduction in cell number but that the cell viability remained unchanged (Nahman et al., '92). The role of the cell cycle in DN was investigated and partially clarified. When cultured under high glucose, HMCL appeared locked in the G1/S transition checkpoint of the cell cycle, and the cells became hypertrophic rather than apoptotic, with increased matrix protein synthesis and the production of pro-fibrogenic mediators like TGFβ1. This cell cycle lock appears to require a glucose dependent alteration of a cyclin dependent kinase inhibitor (CDK inhibitor) $p27^{kip1}$ (Wolf et al., '97, Wolf et al., '98, Wolf and Ziyadeh, '99). Regulation of expression appeared dependent upon translation, rather than transcription, and was inhibited with a TGFβ antibody. Thus, a link between the cell cycle and TGFβ induced mECM expansion has been established which requires further investigation.

1.10.4 SIGNAL TRANSDUCTION

Specific pathways involving intracellular signal transduction have received much interest in DN in the past decade. Mesangial cells display sensitivity to a wide variety of cellular factors. Secondary messengers like cAMP, cGMP and *ras* proteins have attracted attention due to their involvement in ion channels and intracellular calcium mobilisation and sequestration. Similarly, the action of PKC, downstream of the adenylate and guadinylate cyclases, has been extensively studied in MC due to the involvement of PKC in vascular dysfunction and their close relationship to VSMC (Mene et al., '89). There appears to be a pattern of expression, activation and translocation that is associated with HG cultures and *in vivo* DN (Xia et al., '94, Koya et al., '97). The translocation of PKC isoforms around the cellular architecture was followed using immuno-staining and various stresses similar to DM, high ambient glucose and shear forces. Changes in the location and translocation of PKC isoforms were apparent, as was the effect on secondary signal transduction (Amiri and Garcia, '99). For example PKC δ and ε associates with the translocation from the cytosol-to-membrane in NHMC stimulated with ET-1 whereas under high glucose (30mM, 2 days) this translocation is shunted to a cytosol-to-particulate translocation (Glogowski et al., '99). A pivotal role for PKC activation in DN was also suggested by the ability of PKC inhibitors to reduce the expression of various cytokines, most notably VEGF (Cha et al., '00) and TGFβ1 (Studer et al., '95).

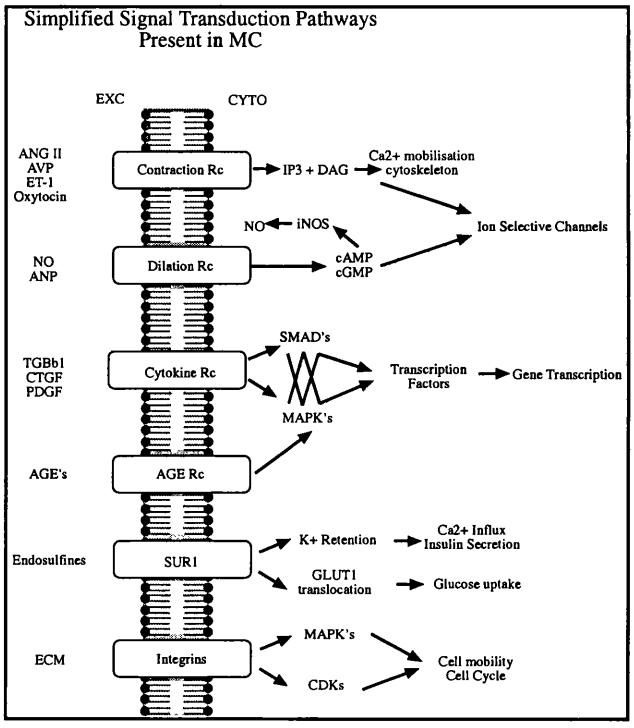


FIGURE 1.9. SIGNAL TRANSDUCTION PATHWAYS PRESENT IN MESANGIAL CELLS.

Irrespective of the origin of the stimuli MC contain many characterised signal transduction pathways. Classic contraction receptors make the MC susceptible to many vasoactive peptides. MC's also possess receptors for many growth factors and ECM receptors (integrins). Detoxification of AGEs through RAGEs is well characterised in MC as is the SUR1 potassium channel. Activation through these pathways can alter many cellular processes including gene transcription.

The TGFβ family of cytokines transduce signal via the SMAD proteins. This particular set of secondary messengers has also been studied in cultured MC. In concert and interwoven through the SMAD and PKC family of transducers are the mitogen activated protein kinases (MAPKs), a family of kinases that take part in a variety of cellular processes also present in the MC (Tomlinson, '99) (see FIGURE 1.9). The result of these signal transduction pathways can be multifactorial. There are many reports of these signal transduction pathways initiating or enhancing gene transcription, the cellular location of PKC isoforms, progression through the cell cycle and apoptotic activation. The eventual elucidation will be complex.

1.10.5 METABOLIC MECHANISMS

Metabolic disturbances in the MC or *in vivo* models of DN have been reported to function through three main mechanisms, all of which were discussed previously in this chapter. The formation and accumulation of AGEs appears a major consequence in the metabolic disturbances in MC. The AGE can take the form of persisting matrix proteins, for example modified collagens that are not cleared through the normal turnover, or in the inducers of intracellular signals through cell surface receptors for AGEs. The most studied of these is the RAGE (receptor for AGEs) but six other binding proteins exist for AGEs (Neeper et al., '92, Li et al., '96). The accumulation of sorbitol through the increased flux of glucose through the polyol pathway was supported through the observation that ARIs reduced the activation of PKC and TGFβ1 in MC grown under high glucose (Ishii et al., '98). The hexosamine pathway is also present and contributes to MC dysfunction. Increased flux through the hexosamine pathway is strongly associated with altered PKC activation and TGFβ1 induction *in vitro* and *in vivo.* Sequentially enhancing and then blockading GFAT activity is sufficient to increase then decrease the activation of TGFβ1 and the over production of matrix proteins (Schleicher and Weigert, '00, James et al., '00). Oxidative stress is recognised as a significant contributor to the diabetic state and it's apparent mechanisms are present in cultured MC. It remains unclear whether oxidative stress is a primary causative mechanism or simply a result of other primary mechanisms (Baynes and Thorpe, '99, Lehmann and Schleicher, '00).

The study of glucose transporters in MC has also been investigated, where it was shown that over expressing GLUT1 was sufficient to increase uptake of extracellular glucose and subsequently increased fibronectin synthesis (Heilig et al., '95). It has also been described that glucose is sufficient to increase expression of GLUT1 and that this leads to increased AR expression and PKCα activation (Heilig et al., '97b, Henry et al., '99).

1.10.6 MECHANICAL STRAIN

Rigidity within the glomerulus is an important factor in haemodynamics, which is regulated in part by the mesangium. In DN the expansion of the mECM is closely linked to changes in intra-glomerular pressure and much work is currently involved in determining whether mechanical strain causes mECM expansion or whether mECM results in mechanical strain (Harris et al., '92, Harris et al., '94, Riser et al., '96). When MC cultures are subjected to shear forces *in vitro* they begin to secrete matrix components. However, when MCs are cultured under high glucose they also begin to produce matrix components. Examining the combination of high glucose and mechanical strain has demonstrated an altered turnover of collagen, resulting in a net increase. It would appear that high glucose serves to aggravate the processes of matrix turnover in the setting of hypertension (Cortes et al., '97).

1.10.7 TRANSCRIPTION OF GENES

Because the transcription of genes has been described in glucose stressed MC, the study of transcription factors and glucose response elements have been investigated and a large amount of data has been published regarding the transcription of genes in response to glucose.

A number of studies have examined the expression of specific genes and proteins expressed in mesangial cells in response to high concentrations of glucose (see TABLE 1.6). These have described the induction of genes, over various time scales, associated with the mesangial matrix, such as fibronectin, laminin and type IV collagen (Haneda et al., '91, Ayo et al., '91a, Ayo et al., '90, Holmes et al., '97, Page et al., '97, Murphy et al., '99). The selection of candidate genes is straightforward but dependent on the rationalisation of their possible role in the cellular response. For example, as

discussed previously, the MC has a disrupted cell cycle during hyperglycaemic stress, thus determining the role of the cell cycle should reveal targets that are modulated by glucose.

Gene	Transcription change	Time Scale, Model	Reference
c-fos/c-jun	2 fold	15min-48h	(Kreisberg et al., '94, Wilmer and Cosio, '98)
Ingrinβ1 (FN-1rc)	Increased	24hrs	(Ihm et al., '95)
$P27^{kip}$	Increased Protein	HMCL-0-96hrs	(Wolf et al., '97)
CTGF, PAI-1, FN1	3fold increase	NHMC-21 days	(Murphy et al., '99)
TGFβ	2fold	48h	(Hoffman et al., '98)
h gremlin	2fold increase	21 days	(McMahon et al., '00)
FN-1, Laminin, Type IV Collagen	2fold increase	0-96hrs Rattus sp. MC	(Ayo et al., '91b)
P4Halpha	7.9	21 days	(Holmes et al., '97)
TSP-1	2.3		
EST (ZnF like)	Significant increase		
HGRG-14	Significant decrease	21 days	(Abdel Wahab et al., '98)
FN1	No change	7days	(Nahman et al., '92)
PCNA, TNF-α, PDGF-B, TGFβ1, bFGF	1.4-5.2 fold increases	2-24 weeks	(Nakamura et al., '93)
IGF-1, PDGF-A, EGF	No changes		
IGFbp2	-3.1	24hrs MMC(SV40)	(Horney et al., '98)
TGFβrcII	2-2.6 1.1-1.4	*In vivo*1-6 weeks *In vitro* 24-72hrs	(Isono et al., '00)
TGFβrc1/II	1.4-1.5	24-48hrs	(Riser et al., '99)
TGFβ1, Biglycan	1.2-2.5	STZ Mus 1-15wk	(Yamamoto et al., '93)

TABLE 1.6. GENE TRANSCRIPTION IN RESPONSE TO GLUCOSE OR IN DM.

The study of differential transcription of genes in response to glucose is currently unearthing many factors that are involved in DN. The variety of model systems and time scales means that a confusing picture of glucose induced differential transcription is being generated. Currently it is unclear which initialling process is involved in the transcription of genes and indeed which genes are necessary and sufficient for DN. The most prominent genes are TGFβ1, a pro-fibrotic cytokine with positive and negative mitogenic properties and CTGF, a growth factor associated with matrix expansion.

Similarly, as expansion of mECM is the hallmark of DN, determining the difference in ECM production or inducers of ECM production should reveal matrix proteins that are modulated by hyperglycaemia. An additional theory regards the changes in contraction seen in DN. Investigating the mechanism of transduction of

stretch stimuli could reveal why this function is impaired in DN. Conjectural formulation like candidate gene selection is an established method for investigation and hypothesis formulation, yet in a complicated system like transcription will only lead to a fragmented picture.

It is likely that several mechanisms may cause altered gene transcription (illustrated in FIGURE 1.10). While a handful of genes have been investigated based on empirical evidence, it is reasonable to expect there are other, un-described genes, which are targets for transcription. These genes may also contribute to the initiation and progression of diabetic nephropathy through either a novel mechanism or by contributing to mechanisms currently understood. A more advanced understanding of the transcription of genes during the early stages of glucose stress may clarify a particular mechanism in DN.

1.10.8 CURRENT HYPOTHESIS OF MESANGIAL CELL CONTRIBUTION TO DN

From observations *in vivo* and *in vitro* it would appear the mesangial cells are target cells in the pathology of diabetic nephropathy and that the primary causative agent is glucose. This hypothesis states that glucose has the ability to alter the mesangial cell in such a way as to reduce the capacity to fulfil contractile function, cause a net increase in the mECM and contribute to the abnormal production or altered sensitivity to cellular factors (see FIGURE 1.10). The physiological result of this glucose stress is the loss of glomerular function. The *in vitro* culture of both primary and transformed mesangial cells has become a standard model for the study of mesangial involvement in DN. It has been established that irrespective of the proliferative state of cultured glomerular mesangial cells, culturing them under high concentrations of d-glucose is sufficient to cause a state of imbalance that will result in cellular responses similar to those associated with DN and seen *in vivo.* This imbalance has been described with respect to the metabolic state of the cell as well as the transcriptional state. The complex nature of DM suggests that other factors exist that are closely involved in the development and progression of DN. Such factors include the paracrine action of cellular factors such vasoactive agents or structural factors like AGEs. However, the transcription of genes plays an important function in the regulation of the

MC and it is likely that there are undescribed genes with regard to the MC *per se* as well as glucose induced alteration of gene transcription.

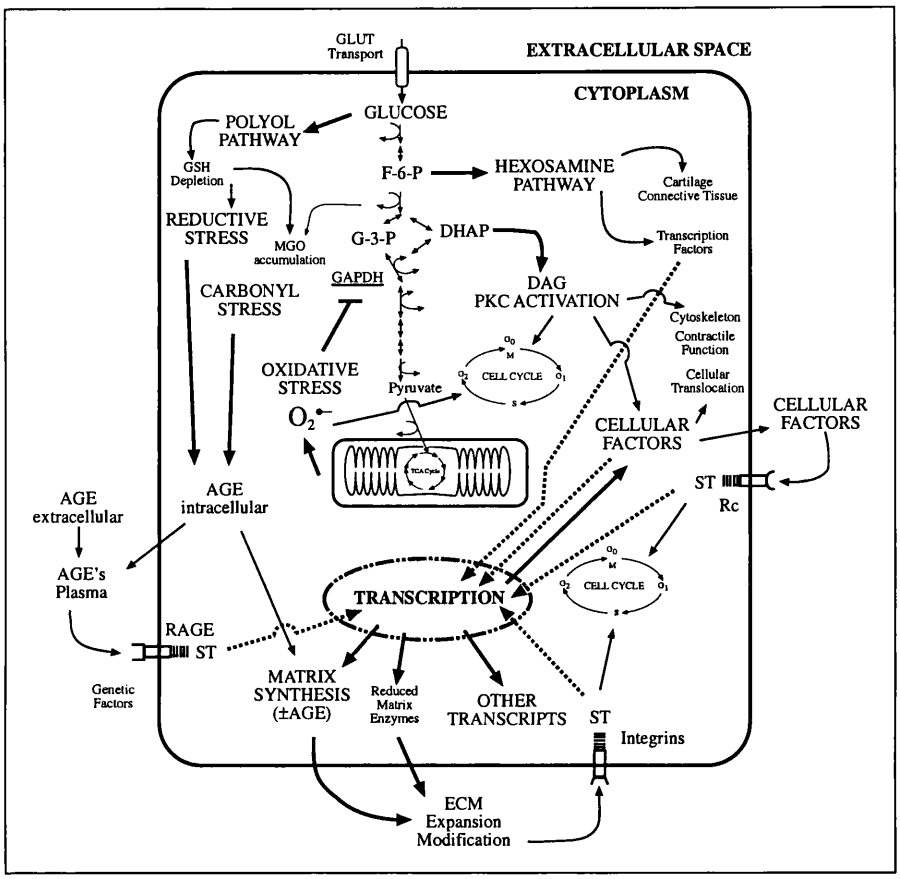


FIGURE 1.10. SUMMARY OF THE PATHWAYS IMPLICATED IN DN.

High glucose can result in a hyper-metabolic state characterised by over activity in glycolysis, the polyol pathway, hexosamine pathway, PKC activation as well as increased production of AGEs. All pathways have the ability to alter transcription of genes either directly or indirectly (dashed lines), which can then act in an autocrine manner where the MC is stimulated or in a paracrine manner where other cells of the glomerulus are implicated. The genes that are altered in response to this glucose stress include matrix factors of the mesangial ECM and mitogenic factors like TGFβ1 and CTGF. It is likely that there are other, un-described factors that are transcribed in response to glucose (ST = Signal Transducation).

1.11 HYPOTHESES UNDERLYING THIS THESIS

The experimentally derived SAGE transcriptome can catalogue the genes transcribed in a cell, describe its phenotype and predict its function. Culturing NHMC under high glucose concentrations alters the transcription of genes and is an *in vitro* model of diabetic nephropathy. Un-described genes exist in cultured mesangial cells that can be identified in an experimentally derived transcriptome. These genes will have pathological significance in DN

1.12 AIMS OF THIS THESIS

1. Construct a SAGE library of genes expressed in Normal Human Mesangial cells.
2. Contribute to the expanding SAGE mapping information with regard to previously un-described genes in NHMCs.
3. Identify and describe genes that are differentially transcribed in NHMCs in response to D-glucose

CHAPTER 2

2 METHODS

2.1 General Overview of Technical Protocols

Unless otherwise indicated all biochemicals used in these experiments were obtained from SIGMA-ALDRICH. All cell culture reagents were obtained from Bio Whittaker UK (Clonetics) with the exception of FCS (Gibco-Life Technologies). All DNA and RNA solutions were, unless otherwise indicated dissolved in TE (10mM Tris HCl, 1mM EDTA, pH8.0). The complex nature of a SAGE analysis meant that the complete protocol has been described in this chapter rather than any of the results chapters. Many of the commonly used reagents are described in ABBREVIATIONS or where specific solutions are used, they are described in the text.

2.2 Cell Culture

2.2.1 THP-1

THP-1 cells (ATCC No.YIB-202) are a monocytic cell line derived from an individual with acute monocytic leukaemia (Tsuchiya et al., '80). The cells were recovered from cryostorage and grown in continuous suspension in RPMI 1640 supplemented with 10%FCS, 1% L-glutamine, penicillin/streptomycin (10,000u/ml, 10,000μg/ml) and 20μM β-mercaptoethanol. Cell suspensions were grown through a 7day cycle in 5% CO_2 at 37°C in a humidified incubator. Cells were harvested by centrifugation at 1400rpm (400×g) for 5min then re-suspended and split 1/5 for continued growth. When required the cell suspensions 1×10^5 cells/ml were stimulated with PMA (160nM). During the subsequent 24 hrs the cells began to fall from suspension and attach to the culture vessel and start to behave as mature macrophages (Tsuchiya et al., '82). Culture media, supplemented with 160nM PMA was replaced every 48hrs for 7days. The cells were then considered fully differentiated into a macrophage-like phenotype, were harvested for total RNA as described below.

2.2.2 NHMC

Primary Human mesangial cell cultures (Bio Whittaker UK Ltd) were established in defined proprietary medium (5.6mM D-glucose, trace Insulin) as

instructed by the distributor and cultured in the presence of 5% FCS. Cells (500,000) were obtained cryo-preserved at passage 3 and were recovered into 6 T25 flasks, then grown through 1 further passage. When the cells had reached 80-90% confluence the cells were harvested using trypsin/EDTA (described below) and cells cryo-preserved for future use in establishing independent cultures. In order that there were sufficient cells to generate the mRNA quantities required for SAGE, cells were grown through 3-4 further passages making the maximum passage of 8 for all the culture series. The cells had not begun to display suppressed growth, generally observed after the 8th passage. Transformed mesangial cells were available both in the laboratory and through commercial means though it was considered that primary human mesangial cells would provide a more physiological significance than their transformed counterparts. Also, as SAGE analysis is such a high-resolution transcriptional analysis the presence of transformation elements was considered undesirable.

Cells (500,000 cell aliquots) were recovered from liquid nitrogen and split into T25 culture flasks (3,500cells/cm^2), containing 5mls of growth medium (MsGM-MsBM, 5%FCS, 50μg/ml gentamicin, 50ng/ml amphotericin-B). The flasks were incubated at 37°C, 5% CO_2 in humidified incubators. Growth medium was changed 24hrs after recovery to remove excess DMSO and then every 48hrs until 90% confluent, generally observed to be 5 days. Population doubling was observed to be between 30-40 hrs.

To sub-culture, the mono-layers were washed twice with 0.15ml/cm^2 HBSS (Hanks Buffered Saline Solution) then treated with 0.12ml/cm^2 trypsin/EDTA (0.025%trypsin/0.01%EDTA) at R/T for 3min (but no more than 7min) or until the cells become detached. Trypsin was neutralized with trypsin neutralising solution (TNS, Bio Whittaker) and cells collected into a sterile 20ml centrifuge tube. Cell recovery and viability was assessed by trypan blue exclusion and cells were concentrated by centrifugation at 1000rpm (228×g) for 5min then re-suspended in MsGM. Cells were seeded into fresh flasks (3,500cells/cm^2) and returned to the incubator. The growth medium was replaced 24hrs after sub-culturing to remove excess trypsin and then every 48 hrs until 90% confluent.

When sufficient cell number was attained, the final subculture was seeded into 34 T25 flasks and/or 10 T175 flasks at 7,000 cells/cm^2 seeding density. Prior to

exposure to elevated levels of D-glucose the cells, 50-60% confluent, were made quiescent by the stepwise removal of FCS from the culture medium: 24hrs in 1%FCS/0.16%BSA (fatty acid free) followed by 24hrs in 0% FCS/0.2%BSA. After this point all culture media contained no FCS and was supplemented with BSA (0.2%). Throughout the time course the cells remained firmly attached to the culture vessel and exhibited minimal proliferation (personal observation). A time scale was initiated at t = -48hr hr and 24hr samples were removed through to t = +120hr. These included culture media, used to test the D-glucose level and total RNA (TRNA). At t = 0 culture medium was replaced with fresh media that contained 5.6mM D-Glucose, 30mM D-glucose or 5.6mM D-glucose +24.4mM mannitol (equivalent osmotic stress). Growth medium was replaced every 48hr for the entire experiment. For t = -48, -24, 0hr two T25 flasks (1.2×10^6 cells each point) were processed for TRNA extraction (see below). For t = 24, 48, 72, 96 and 120 hr six T75 flasks (1.2×10^6 cells each point) were processed for TRNA extraction. At t = 48hr and t = 96hr, 4x T175 flasks (8.2×10^6 cells) were processed for mRNA isolation and SAGE (2 for high D-glucose and 2 for low D-glucose). Samples were stored at -70°C until required. D-glucose concentration remained constant throughout the experiment as determined by the Unimat5 D-glucose hexokinase Kit (ABX). The entire sequence from recovery from liquid nitrogen to t = 120hrs was generally 26 days. The cell culture sequence was repeated to create a triplicate set of RNA sample with which to construct the SAGE libraries, and a further four culture series were undertaken later in the project, for use in corroboration experiments.

2.2.3 HMCL

Transformed human mesangial cells (Sraer et al., '96) were cultured in $DMEM_{1000}$ (5.4mM D-glucose, Life Technologies), supplemented with 5%FCS, 2mM L-glutamine and antibiotics (50μg/ml gentamicin, 50ng/ml amphotericin-B), in humidified incubators at 37°C and 5% CO_2. Sub-culturing was conducted as above and cells were seeded at 4000cells/cm^2 for each passage and 8000cells/cm^2 for each time sequence. Sub-confluent cells were rendered quiescent by the stepwise removal of serum over a 48hr period (1% FCS/0.16%BSA 24hrs followed by 0 %FCS/0.2% BSA 24hrs). Cells were then divided into three groups and cultured for a further 120hrs in serum free media containing 5.4 mM D-glucose, 30mM D-glucose or 5.4mM D-glucose plus 24.4mM mannitol (equivalent osmotic stress). Media was replaced every 24hrs to

ensure constant D-glucose levels. Starting from t = -48hrs (proliferating mesangial cells), triplicate T25 flasks were processed every 24hrs for total RNA isolation. All processed samples were stored at -70°C until required.

2.2.4 CRYO-PRESERVATION OF CELLS

Cells were stored long term in liquid nitrogen. Cryo-preservation was carried out according to a protocol obtained from the distributor (Bio Whittaker UK). Briefly, cells were grown to 80-90% confluence, or to between 1-3×10^6 cells/ml for suspensions, and the cells were collected by the trypsin/EDTA method, described above, or centrifugation. The cells were re-suspended to a concentration of 1×10^6 cells/ml in EMEM supplemented with 20%FCS. The cells were then diluted further with an equal volume of EMEM (15%DMSO), to a final concentration of 10%FCS and 7.5%DMSO. The cells were not allowed to spend more than one hour at R/T before initiating the freezing protocol. The cells were aliquoted into cryovials (Corning UK) such that a single vial contained approx 500,000 cells and then placed in an alcohol bath at R/T (1-propanol). The container was placed at –80°C overnight. Once frozen (at least 4hrs at –80°C) the cells were then transferred to liquid nitrogen for long-term storage.

When required, the cells were recovered from the liquid nitrogen by rapidly thawing them at 37°C and distributing them between 5 T25 flasks containing 5ml of pre-warmed MsGM, at an approximate seeding density of 3500 cells/cm^2. After 24hrs incubation at 37°C, 5%CO_2 the media was changed and the cells incubated for a further 24hrs. Generally, the cells would adhere to the flask and begin to proliferate after 24hrs. The efficiency of reconstitution was approx 80%, as determined by crude cell counting under low power microscopy.

2.3 RNA ISOLATION

2.3.1 ISOLATION OF TOTAL RNA FROM CELL SUSPENSIONS

Cell suspensions (10 ml) were collected in 15ml centrifuge tubes (≅ 100,000 cells/ml) and the cells collected by centrifugation at 1000rpm (220×g) for 5 min. The

supernatant was removed and the cell re-suspended in 10ml of warm PBS. The cells were again centrifuged and washed in the same manner. After this final wash, the cells were re-suspended in 500μl of PBS and then lysed by the addition of 5ml of RNA Isolator (Genosys Biotechnologies). The lysed cells were mixed gently at R/T for 5 min and repeatedly drawn through a 19G needle. Aliquots of 1ml were distributed to 2.0 ml tubes and stored at –70°C until required. Total RNA (TRNA) was isolated from the thawed lysate by the addition of 0.2ml of chloroform followed by vigorous shaking. The phases were separated by centrifugation at 12,000 rpm (13,400×g) for 10min at 4°C. The upper aqueous phase was removed to a clean 1.5ml tube, being careful not to disturb the lower phenol or inter-phase, and the TRNA precipitated with 0.5ml of 1-propanol. TRNA precipitation was allowed to proceed for 15min at R/T then collected by centrifugation for 10min at 12,000 rpm (13,400×g). Pellets of TRNA were washed with 1ml of 75% ethanol then air dried for 5 min. The RNA was dissolved in RNAse-free TE or water and the concentration and purity determined by measuring the A_{260} and A_{280} where 40μg/ml RNA is equivalent to 1 OD_{260} and an A_{260}/A_{280} value between 1.8 and 2.0. TRNA samples were stored at -70°C until required.

2.3.2 Isolation of Total RNA from Cell Mono-Layers

Monolayers were washed twice with warm (37°C) PBS then aspirated to remove as much liquid as possible from the culture flask. The cells were lysed directly in the flask by the addition of 1ml/25cm^2 of RNA Isolator reagent (Genosys Biotechnologies) and incubated at R/T for 5min. The lysate was passed several times through a 19G needle and syringe and 1ml aliquots were stored in 2.0ml tubes at -70°C until required. Total RNA was isolated as described above.

2.3.3 Purification of mRNA from TRNA

Total RNA (80μg) from cells grown under high or normal glucose was prepared from RNA Isolator as above and re-suspended in 100μl TE. Further purification to isolate mRNA was accomplished by binding to oligo-dT cellulose using the MicroFast Kit (Invitrogen).

The TRNA sample was adjusted to 900μl using dilution buffer (200mM NaCl, 200mM Tris pH7.5, 1.5mM $MgCl_2$ and 2% SDS), heated to 65°C for 10min then quenched on ice for 5min. The NaCl concentration was adjusted to 0.5M by adding 63μl of 5M NaCl.

A 25mg aliquot of oligo-dT cellulose was added to each of the samples and incubated for 2min at RT to allow the oligo-dT cellulose to swell. The samples were rocked gently at R/T for 45min to allow the polyA$^+$ RNA to anneal to the oligo-dT cellulose. Annealed polyA$^+$ RNA was collected at 6,500rpm (4,000×g) for 5min and the supernatant carefully removed. The resin was washed 3 times at R/T with 1.3ml of binding buffer (500mM NaCl, 10mM Tris pH 7.5). Following each wash the resin was collected by centrifugation at 4,000×g for 5min. The resin was then transferred to a spin column and washed with 500μl of binding buffer by spinning at 6,500rpm (4,000×g) for 10s. The eluate was collected and A_{260} read. The resin was washed repeatedly in binding buffer until the A_{260} of the eluate was < 0.05. The resin was then washed twice in a similar manner with 200μl of low salt buffer (250mM NaCl, 10mM Tris pH7.5). The resin was re-suspended in 100μl of elution buffer (10mM Tris pH7.5, prewarmed to 50°C) and the mRNA eluted by centrifuging for 1min at 6,500rpm (4,000×g). This step was repeated with an extra 100μl and the combined eluate (200μl) transferred to a clean tube.

The mRNA was precipitated by the addition of 10μl of 2mg/ml glycogen (Boehringer Mannheim) as a carrier, 30μl of 2M NaOAc and 600μl of 100%ethanol. The precipitation was stored at -70°C until required. The yield of mRNA was calculated by reading the A_{260}/A_{280}.

2.3.4 PREPARATION OF cDNA

PolyA$^+$ RNA was converted to double stranded cDNA by the modified method of Gubler-Hoffman using the CopyKit (Invitrogen) with the substitution of a 5'-biotin labelled polyT primer. Purified mRNA (3μg) was primed by mixing 1μg (80pmol) 5'-biotin-T_{20}-3' primer in 30μl of dH_20, heated to 65°C for 5min then allowed to cool to RT. The reaction buffer was assembled (0.2U/μl RNAse Inhibitor, 100mMTris-HCl, 40mM KCl, 10mM $MgCl_2$, 0.5mM spermidine, 4.0mM dNTPs, 4.0mM $Na_2H_2P_2O_7$,

pH8.3). First-strand synthesis proceeded with AMV reverse transcriptase (1U/μl) at 42°C for 1hr. Completed first strand reaction was chilled on ice for 5min. Incubation buffer was adjusted (160mM KCl, 50μg/ml BSA, 0.2mM β-NAD, 10mM DTT). The DNA:RNA hybrid was treated sequentially with *E.coli* RNAse H (0.012U/μl), *E.coli* DNA Polymerase (0.1U/μl) and DNA Ligase (0.1U/μl) and incubated at 16°C for 90min, then 30min at R/T. Heating the reaction to 70°C for 10min then chilling on ice halted second strand synthesis. A final step treated the cDNA with T4 DNA polymerase (0.012U/μl) for 10min at RT.

Blunt-ended, 3'biotin labelled cDNA was extracted with an equal volume of P/IAC and precipitated with the addition of glycogen (to 40μg/ml) and NH_4OAc (to 2.0M) then 2vol of 100% ethanol. The precipitated cDNA was collected by centrifugation at 12,000rpm (13,400×g) for 20min and washed twice with 500μl of 80% ethanol. The air-dried cDNA was re-suspended in 30μl of TE and the A_{260} and A_{280} were used to determine concentration. cDNA was stored at –20°C until required.

2.4 HYBRIDISATION EXPERIMENTS

2.4.1 HYBRIDISATION TO GENEFILTERS™

GeneFilters™ were hybridised as directed by the manufacturer and analysed using the Pathways™ 2.01 software (Research Genetics). Prior to the first hybridisation the GeneFilters were washed for 30min in boiling 0.5%SDS. The filters were placed into a 50ml polypropylene tube to which was added 5ml of pre-warmed MicroHyb solution (Research Genetics) that had been supplemented with 5μg of denatured Cot1 DNA (non-homologous DNA enriched for repetitive sequences) and 5μg polyA oligo nucleotides. The pre-hybridisation was allowed to proceed for 2hr at 42°C with continuous revolving in a hybridisation oven (Techne).

During pre-hybridisation, probe synthesis and clean up was carried out. Briefly, total RNA (2.0μg) was primed with oligo-dT (2μg) by heating to 70°C for 10min followed by chilling on ice for 2 min. The solution was adjusted with reverse transcription buffer 50mM Tris HCl 8.3, 75mM KCl, 3mM $MgCl_2$, 3mM DTT, 0.05mM each dGTP dATP and dTTP. Reverse transcription was initiated with the addition of

20U SuperScript™ (Life Technologies) 10U/μl and 10μl of α-^{33}P-dCTP (2500Ci/mmol) and incubated at 37°C for 90min. Unincorporated radiolabel was removed from the reaction by chromatography through a BioGel P60 acrylamide resin (BioRad) or Sephadex G25 (Amersham-Pharmacia biotech) columns equilibrated with TE. The column was drained by centrifugation at 1750rpm (700×g) for 4min. The labelling reaction was diluted to 100μl and applied to the column, which was spun again for 4min at 1750rpm (700×g). The eluate containing the probe was denatured by heating to 96°C for 5min, then chilled on ice prior to adding to the pre-hybridised GeneFilter and allowed to hybridise for 12hr at 42°C. The filter was washed twice in 2x SSC (0.3M NaCl, 0.03M NaCitrate), 1% SDS at 50°C for 30 min and once in 0.5×SSC 1%SDS at 55°C for 20min.

The filter was exposed to a low energy phosphor storage screen (Kodak) for 48hrs and scanned using the maximum resolution on a Storm™ 860 Imager (Molecular Dynamics). The image was cropped in the ImageQuant™ software, and imported as single filters directly into the Pathways™ 2.01software and analysed. It was important not to alter the image in any other graphics software as this would result in the permanent loss of data. Prior to repeating the hybridisation with independent RNA, the filter was stripped by agitating in boiling 0.5% SDS for 30min, then re-probed with labelled RNA from the time points t = LG96hr, HG96hr, MT96hr. By this point, the filter had suffered too much loss of sensitivity to consider further hybridisation.

2.4.2 HYBRIDISATION TO NORTHERN BLOTS

Northern analysis was accomplished using standard techniques. Total RNA (5μg), was separated using 1.5% denaturing MOPS/Formaldehyde agarose gel electrophoresis. Briefly, 1.5g of agarose was dissolved in 73ml of water and 10ml of 10× MOPS buffer (0.4M MOPS (pH 7.0), 0.1M NaOAc, 10mM EDTA). Once cooled to 50°C, 17ml of 37% formaldehyde was added, mixed and then the entire solution was cast in a gel tray and allowed to solidify for 1 hr. The gel was transferred to an electrophoresis tank filled with 1× MOPS buffer and pre-run for 30min at 15V/cm. Meanwhile the TRNA was heated to 65°C for 10 min then cooled on ice for 5min. To the RNA was added, 2vol of formamide, MOPS buffer to 1× and formaldehyde to 20%. The cooled RNA solution was loaded onto the gel and electrophoresed for 2hr at 15V/cm.

Capillary blots were assembled using 20×SSPE (3.6M NaCl, 0.2M Sodium phosphate, 0.02M EDTA pH 7.7) as the transfer solution and the separated RNA was transferred overnight onto Hybond N^+ nylon membrane (Amersham Biosciences). Following transfer, the blotting stack was dismantled and the filter washed briefly in 2×SSPE and allowed to air dry for 10 min. The RNA was cross-linked to the filter by exposure to UV light for 2min. The filters were stored dry at R/T until required.

Probe fragments were identified in nucleic acid databases and entire clone inserts were PCR amplified from IMAGE clones (see 'PCR Amplification of vector insert DNA'). Probes for P4HP, TSP1 and GLUT1 were constructed using purified inserts from IMAGE clones identified through UniGene (ncbi.nlm.nih.gov/unigene) as clustering with a characterised cDNA of the gene of interest. Gene fragments were labelled with ^{32}P and hybridised using standard techniques. GAPDH signals from stripped and re-probed filters were used to normalize the signals. Time points from t = 0hr through to t = 120hr were examined from these experiments. These fragments were quantified and 25ng of probe template was used in a random primed labelling reaction. The probe template was denatured at 95°C for 5min, and then quenched on ice for 5min. The template was added to a Rediprime™ II random priming reaction tube containing components for random primed DNA synthesis, random nonomers, dNTPs, α-^{32}P-dCTP (3000Ci/mmol), and Klenow (DNA polymerase) (AmershamBiosciences). The labelling reaction was allowed to proceed for 1hr and the unincorporated label separated by gel chromatography through 1ml of Sephadex G-50 resin. The labelled probe was eluted in 400μl of TE and denatured by heating to 95°C for 5min then quenched on ice.

Filters were pre-hybridised in roller bottles (Techne) in 10ml of pre-hybridisation solution (5×SSPE, 5×Denhardt's and 0.5%SDS, 20μg/ml non-homologous DNA) for 1hr at 65°C. The denatured probe was added directly to this pre-hybridisation solution and allowed to hybridise to the filter for 12hr. Once complete the filters were washed twice for 30min at R/T in 2×SSPE, 0.1% SDS and then twice for 30min at 60°C in 0.5×SSPE, 0.1%SDS. The filters were then wrapped in Saran film and exposed to X-Ray film (Kodak X-OMAT AR) for 3-6hr.

2.4.3 HYBRIDISATION TO DOT BLOTS

IMAGE clone gene fragments (1μl) (see 'IMAGE clones' below) were spotted in serial dilution (2.0, 1.0, 0.5 nM) on Hybond N^+ membranes (AmershamBiosciences), air dried for 20min then exposed to UV (320nm) for 2 min. The filters were stored at R/T until required. Each filter was hybridised to α-^{32}P-dCTP labelled ssDNA that was reverse transcribed from 2μg total RNA in the same manner as the probe synthesis for GeneFilters™. Hybridisation proceeded at 60°C in 5ml of ExpressHyb (CloneTech) for 12hrs. Filters were washed at R/T twice in 2×SSC, 0.5%SDS for 30 min then twice at 60°C in 0.5×SSC, 0.5%SDS for 30min. The filters were exposed to both auto-radiographic film (Kodak X-OMAT AR) and a phosphor storage screen (Molecular Dynamics). Point densities were determined using Quantity One (BioRad) or ImageQuant (Molecular Dynamics).

2.4.4 DETERMINING BAND OR SPOT DENSITY

Images on autoradiographic film were scanned in to a high definition TIFF file for maximum resolution. These files were uploaded in the QuantityOne software package (BioRad) or, in the case of the phosphor screens ImageQuant (Molecular dynamics). Saturation was avoided and the global background was removed. Spot or band density was determined with all calculation areas constant to simplify the normalisation process. Density values were exported into MS Excel spreadsheets and mathematically manipulated to normalise each density value to a set of reference densities, generally the genomic DNA to adjust for filter-to-filter variation and housekeeping genes to adjust for RNA sample variation.

2.5 IMAGE CLONES

IMAGE clones (Integrated Molecular Analysis of Genome Expression) are a collection of some 4.1 million distinct cDNA clones that are maintained and distributed through the Human Genome Mapping Project (HGMP) to researchers worldwide. This collection is sourced from some 450 cDNA libraries representing more than 48 tissue types and/or stages of development. The primary motivation of this IMAGE consortium is to assist in the discovery and classification of expressed genes (Lennon et al., '96).

2.5.1 IDENTIFYING IMAGE CLONES

An experimental clustering system has been in operation at the NCBI for some time now. UniGene partitions Genebank and EST sequences into clusters that suggest a unique gene. UniGene clusters contain links to other databases, such as the 'Online Medelian Inheritance in Man', (OMIM), which compiles gene description and genetic traits, LocusLink, which maps genes to genomic locations and the various nucleic acid databases that contain the sequences of the characterised mRNAs, ESTs and STSs that form the UniGene cluster. UniGene clusters however, do not contain consensus alignments or contigs, and the clustering procedure, while repeated regularly in 'builds', is still experimental and subtle alterations in clustering are regularly observed (Schuler et al., '96).

SAGE tag mapping information from NCBI (ncbi.nlm.nih.gov/unigene) and output data from Pathways™ contains UniGene cluster ID that correspond to either tag of signal position of the GeneFilter. These cluster IDs were used to identify IMAGE consortium clones that could be used as probes from further analysis of candidate genes.

2.5.2 GROWING BACTERIAL CULTURE

The IMAGE clones were obtained from HGMP-UK (Hinxton UK) as an E.*coli* bacterial stab that contained the EST as an insert into a common cloning vector, which would confer ampicillin resistance to the E.*coli* host. These were re-plated on fresh LB/ampicillin agar plates (50μg/ml) and grown overnight at 37°C. Isolated colonies, (generally 5), were picked to 15μl of 1× PCR buffer and the inserts amplified by PCR (see 'PCR Amplification of Vector Insert DNA'). As well as PCR amplified inserts; plasmid stocks were created (see 'Mini-prep Plasmid DNA Isolation'). Glycerol stocks of all IMAGE clones were prepared for long-term storage. Briefly 500μl of the overnight liquid culture was diluted 1:2 with LB/glycerol (30%) and stored at –70°C. When re-growth of the clone was required, a sliver of this mixture was removed with a sterile needle to a fresh LB/Amp agar plate without allowing the entire bacterial suspension to thaw completely. The droplet was dilution streaked and incubated overnight at 37°C. Single colonies were easily obtained from this dilution streak.

2.5.3 MINI-PREP PLASMID DNA ISOLATION

Overnight liquid cultures of the bacterial clones (3.0ml LB/Amp: 50μg/ml) were centrifuged for 5min at 5000rpm (2,300g), the liquid removed and the bacterial pellets re-suspended in 200μl GTE (1% glucose, 25mM Tris-HCl, 10mM EDTA, 5μg/ml RNAseA). The cell membranes and proteins were disrupted and the nucleic acid was denatured by the addition of 200μl of lysis buffer (200mM NaOH, 1% SDS), mixed by several inversions then incubated on ice for 5min. The solution was neutralised by the addition of 300μl potassium acetate (KOAc: 3M K^+, 5M Acetate), mixed by vigorous shaking then incubated on ice for a further 10min. The precipitate was removed by centrifugation at 13,100rpm (16,000g) and the supernatant removed to a clean tube. The clarified supernatant was extracted sequentially with an equal volume of P/IAC then the IAC and the plasmid DNA precipitated with 2 vol of 100% ethanol for 15min. Precipitated plasmid DNA was collected by centrifugation at 13,100rpm (16,000g) for 15min, washed twice with 70% ethanol, air dried for 10 min then re-suspended in 50μl of TE. Plasmid DNA concentration was estimated by reading the UV absorbance at 260nm (A_{260}) and stored at –20°C. If plasmid DNA was required for DNA sequencing then the plasmid preps were further purified by polyethylene glycol (PEG) precipitation. An equal volume of 20% PEG_{8000} in 2.5M NaCl was mixed with the plasmid DNA and the precipitations incubated on ice for 20min then centrifuged at 13,100rpm (16,000×g) for 15min. The plasmid DNA pellet was washed twice with 70% ethanol then re-suspended in 20μl of TE.

2.5.4 IDENTIFYING BY SEQUENCE

Each of the amplified inserts was sequenced to confirm identity (see 'DNA Sequencing'). Briefly, the sequence data, approx 400bp of quality sequence, was entered into a 'BLASTn' rapid alignment query at the NCBI (ncbi.nlm.nih.gov/BLAST). Queries were returned by email with matching identities. Clones were positively identified 76% of the time as being true to cluster ID. Those that failed identification were excluded from the subset and alternatives sought.

2.5.5 SIZE DETERMINATION

Insert size was determined by electrophoresis in a 1.0%TAE-agarose gel with DNA size markers (Lambda/*Hin*dIII fragments, Life Technologies). The gel was stained with ethidium bromide (0.5μg/ml) and visualized with UV trans-illumination. The image was captured by a digital camera (GelDoc, BioRad) and processed with Quantity One software (BioRad). The band sizes (bp) were determined as a function of their mobility when compared to the size markers.

2.5.6 MOLAR CONCENTRATION

In order that there were similar concentrations of target molecules in the hybridisation experiments the relative molar concentration of each IMAGE clone insert was determined using the conversion relationship (EQUATION 2.1). Each purified IMAGE insert was diluted to 2.0nM and stored at –20°C.

$$1\mu g \text{ of } 1000bp \text{ DNA} \equiv 1.52pmol$$ **EQUATION 2.1. MOLAR RELATIONSHIP OF DNA**

For example: $6\mu g$ of 540bp DNA $\equiv 1.52pmol \times 6\mu g/1\mu g \times 1000bp/540bp$

$= 16.89pmol$

2.6 SERIAL ANALYSIS OF GENE EXPRESSION: SAGE

Serial analysis of gene expression (SAGE) is a technique that takes advantage of high throughput cloning and DNA sequencing technology to obtain a quantitative profile of the transcription dynamics. SAGE does not measure the level of a gene, rather the frequency of a 'tag' that represents a transcription product of a gene. Tags are short 10bp lengths of cDNA from a defined position in a transcript that can be serially concatenated, sequenced and mapped with high efficiency to produce a digital representation of the abundance of gene expression.

Three SAGE libraries were constructed for each of the high and low D-glucose cultures series. PolyA$^+$RNA was isolated and taken through the protocols detailed below for each of these series individually before being combined after it was

determined that they contained sufficient similarity with the other two libraries in their group (see FIGURE 2.1).

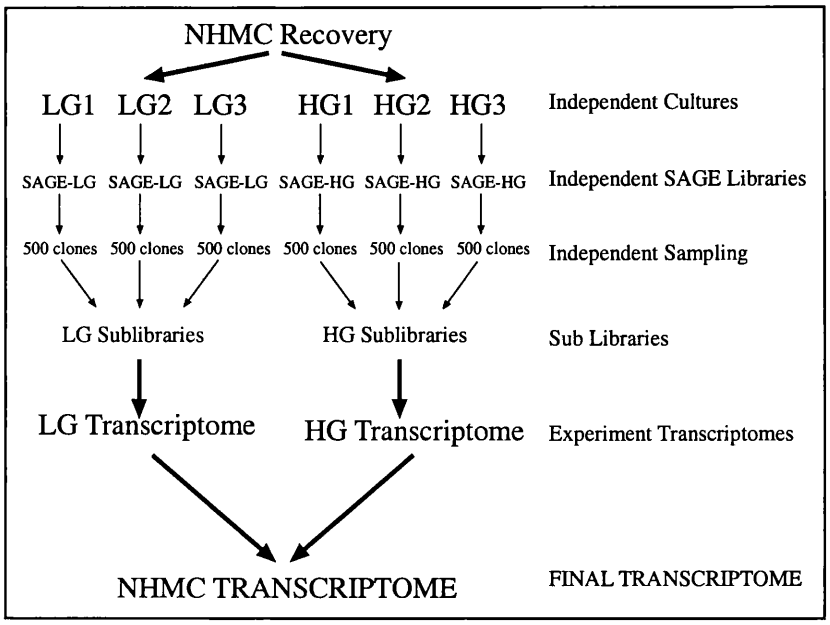


FIGURE 2.1. SCHEMA OF THE CULTURING PROTOCOL.

Independent libraries were produced from triplicate independent cultures. The SAGE libraries were sampled with approximately 500 clones each and the sub-libraries were eventually combined to produce the final NHMC transcriptome. A single culture series takes 26 days and the most time consuming section was the sequencing.

2.6.1 CLEAVING THE CDNA AND BINDING TO DYNAL MAGNETIC BEADS.

A total of 3μg from each cDNA preparation was used to construct the SAGE sub-library. The biotin labelled cDNA was digested in 50μl 1× NEB buffer 4 with 5U/μg N*la*III (10U/μl, New England BioLabs) at 37°C for 2 hr. The digest was then extracted with an equal volume of P/IAC then precipitated with 1/10vol NaOAc and two volumes of absolute ethanol. The digested cDNA was re-suspended in 100μl 0.5×TE. Dynabead M280 Streptavidin (100μl, 10mg/ml) was washed with 100μl of binding and washing buffer (1×BW buffer: 1M NaCl, 1mM EDTA, 5mMTris pH7.5). The slurry was mixed vigorously and then the beads were immobilized using a magnet. This wash was repeated and the beads re-suspended in 100μl 2×BW buffer. The digested cDNA was added directly to this slurry and incubated at R/T for 20min. The captured fragments were washed three times with 200μl of 1×BW buffer then once with 200 μl 0.5×TE.

2.6.2 CREATING SPECIFIC SAGE LINKERS

Linkers were designed that contained ends that are compatible to the anchoring enzyme *Nla*III ends and contained the recognition sequence for the tagging enzyme B*sm*FI (see FIGURE 2.2).

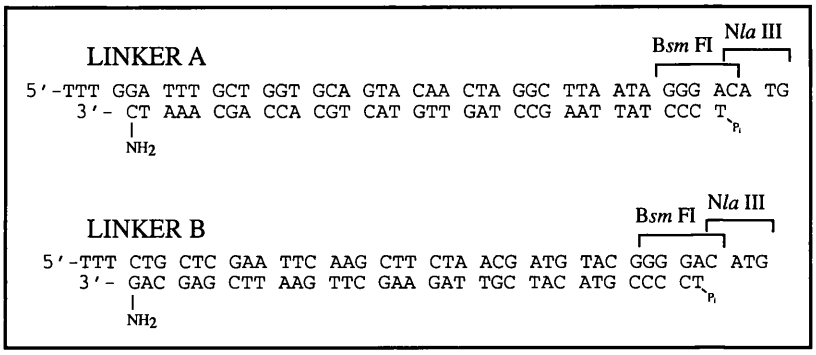


FIGURE 2.2. SAGE LINKERS.

Linkers were designed according to the sequences illustrated. Ligation to N*la* III ends and a B*sm* FI site was engineered into the linker adjacent to the N*la* III site. Amplification primers were designed further 5' to the B*sm* FI.

Individual oligos were purified by denaturing 12% PAGE. Aliquots of the impure oligos were heated to 80°C for 10 min and loaded onto a pre-run 8% polyacrylamide/urea gel (19:1 acrylamide:bis-acrylaminde) and run at 25V/cm for 5 hr. The gel was stained lightly with ethidium bromide (0.5μg/ml) and bands visualized with long wave UV. Excised bands were crushed in 0.5ml of 0.3M NaOAc and heated to 65°C for 10min. The slurry was clarified by centrifugation through a 0.22μm spin filter, extracted with an equal volume of P/IAC, precipitated with 2.5vol of 100% ethanol, washed twice with 75% ethanol and re-suspended in TE at 10nmol/μl.

Equimolar amounts (50nmol) of the oligos were mixed and the salt concentration adjusted to 0.1 M with NaCl. Tubes were then placed in a block heater at 80°C for 10min to denature secondary structure. The block was removed from the heater unit and allowed to cool slowly to R/T, generally around 4 hours. Samples of the now annealed linkers were tested by self-ligation. Briefly, small amounts of linkers were mixed (0.5μg each) in a ligation reaction with 0.1U/μl T4 DNA ligase (1U/μl, Boehringer Mannheim) at 16°C for 2 hr. The ligations, together with un-ligated linkers and a 10bp DNA ladder (Life Technologies) were loaded onto an 8% non-denaturing TAE polyacrylamide gel and run in TAE buffer for 3hr at 10V/cm. The gel was stained with 1×SYBR Green 1 and the DNA bands visualised by UV trans-illumination. Linkers were diluted to reduce NaCl concentration (0.33μM), aliquoted into 0.5ml tubes and stored at –70°C.

2.6.3 LIGATING LINKERS TO THE 5' cDNA AND RELEASING TAGS

The Dynabead slurry containing the captured 3'cDNA ends was re-suspended in 1× ligation buffer (66mM Tris-HCl, 5mM $MgCl_2$, 1.0mM DTT, 1.0mM ATP, pH 7.5) and divided into two 0.5ml tubes. Each linker (1μg or 0.33pmol) was added to one half of the Dynabead slurry containing the cDNA ends, the volume adjusted to 50μl and the ligation initiated by adding 3μl of T4 DNA ligase (5U/μl) then incubating at 16°C for 4hrs with intermittent mixing.

Following ligation the slurry was washed 4 times with 200μl 1×BW and then twice with 100μl NEB buffer 4 (New England BioLabs Restriction Enzyme buffer 4). To release the tags, the washed ligation products were digested with 8U of the tagging enzyme B*sm*FI (4U/μl, New England BioLabs) in 100μl at 60°C for 2hrs with intermittent mixing. After applying the magnet the supernatant, which now contained the released tags, was collected and the beads back extracted with an extra 50μl TE. The combined supernatants were extracted with an equal volume of P/IAC then precipitated with 3 volumes of ethanol. The precipitated tags were washed with 100μl of 75% ethanol, air-dried then re-suspended in 10μl 0.5×TE.

2.6.4 LIGATING THE TAGS TO FORM DITAGS

The 5'overhangs of the released tags were filled to create blunt ends by treatment with E.*coli* DNA Polymerase I Large Fragment (Klenow Polymerase) (1×Klenow buffer, 0.5mM dNTPs, 0.06U/μl Klenow enzyme), with incubation for 30min at 37°C. The reaction volume was increased to 200μl, extracted with an equal volume of P/IAC then precipitated with 3vol of ethanol. The air-dried precipitate is re-suspended in 10μl of 0.5× TE.

To form the ditags the two samples were combined and ligated with 0.7U/μl of T4 DNA ligase at 16°C for 12 hrs. Following incubation the reaction was diluted to 30μl and stored at –20°C until amplification of ditags was required. Then the mixture was diluted further and used to PCR amplify end-to-end ditags.

2.6.5 PCR AMPLIFICATION OF DITAGS

SAGE specific primers designed against the linkers were synthesized (MWG-Biotech) and used in titrated PCR reactions (SAGE primer 1: 5'-GGATTTGCTGGTGCAGTACA-3'; SAGE primer 2: 5'-CTGCTCGAATTCAAGCTTCT-3'). Serial dilutions of the ligation mixture (1μl of a 1/10-1/500 dilution) were added to a 50μl PCR reaction (final template dilution 1/500-1/25000), 1× SAGE-PCR buffer (25mM TAPS, 50mM KCl, 6.7mM $MgCl_2$, 1mM DTT, 1.0mM dNTPs), 1.5μM each primer and 0.5U T*aq* DNA polymerase (Qiagen). The PCR cycling conditions followed the schedule of 30 cycles of 96°C for 30s, 55°C for 10s then 70°C 1min, followed by a final extension of 5min at 70°C.

Products were resolved on a 10% polyacrylamide gel together with 10bp DNA markers (Life Technologies), un-ligated linkers and self-ligated linkers. The gel was stained with SYBR Green 1 (Molecular Probes Inc) and examined with UV trans-illumination. Amplified products ran at approx 100bp and background bands ran at 80bp together with self-ligated linkers. The dilution that presented a strong 100bp product with minimal background was selected to scale up the PCR reactions. Generally this was observed to be a 1/10,000 dilution (FIGURE 2.3a).

a

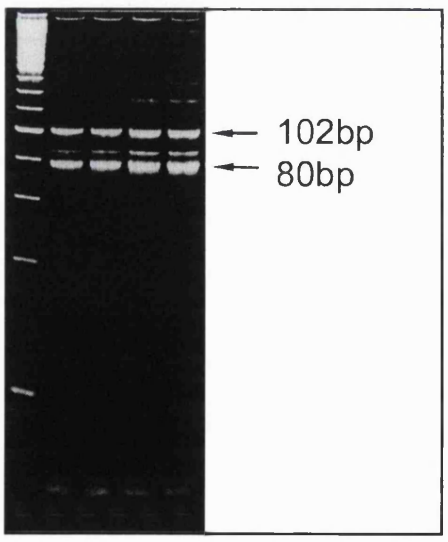
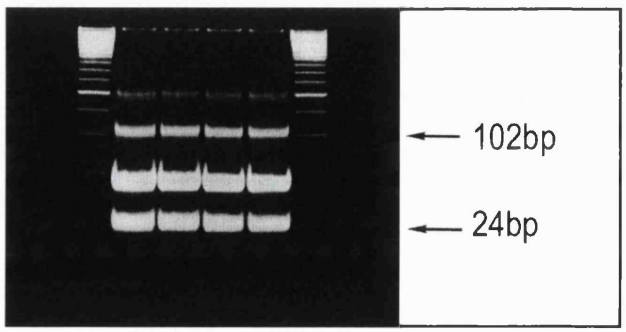


FIGURE 2.3 A & B. PCR AMPLIFICATION OF DITAGS.

PCR amplified ditags appear as a 102bp product. A background band of 80-85bp will generally appear as an equal or lesser intensity band (a). When Amplified ditags are purified and digested with N*la* III the ditag cassette is released from the PCR product (b)

b



Amplification was scaled up to 96 × 100μl PCR reactions. Products from completed PCR reactions were pooled into a 15ml tube and extracted with an equal

volume of P/IAC. Aliquots of 400μl were transferred to clean 1.5ml tubes and precipitated with 2.5 vol of ethanol in the presence of 3M NH_4OAc. Precipitated PCR products were re-suspended in 20μl of TE, pooled into one tube (500μl total) then adjusted to 600μl with 5× loading dye (10%w/v $ficoll_{400}$, 0.025%w/v bromophenol blue, 0.025%w/v xylene cyanol). The products were loaded onto four 10% TAE-PAG (19:1 acrylamide:bis-acrylamide) and run at 5V/cm for 4-5 hr. The gel was stained with 1×SYBR Green 1 for 15min and visualized with long wave UV trans-illumination. The 100bp bands were excised from the gel and placed in sterile 0.5ml tubes. The tubes were pierced, top and bottom, with a 19G needle and placed inside a 2.0ml tube then centrifuged for 2min at 11,400rpm (12,000×g). The fragmented gel slurry was re-suspended in 500μl of 0.5×TE and heated to 65°C for 20min. Filtering through a 0.22μm spin filter removed polyacrylamide fragments and the purified ditags were extracted with an equal volume of P/IAC and precipitated with glycogen carrier (3ng/μl), in the presence of NH_4OAc (2.5M) and 2.5vol ethanol. Ditags were collected by centrifugation at 11,400rpm (12,000×g) for 20min, washed twice with 75% ethanol, then re-suspended in 10μl of 1× NEB buffer 4 (50mM KOAc, 20mM Tris-OAc, 10mM MgOAc, 1mMDTT, pH 7.9) and pooled. The purified ditags were digested with N*la*III (1U/μl) at 37°C for 2hrs. The digestion was adjusted to 200μl and extracted with an equal volume of P/IAC then precipitated with 3vol of 100% ethanol.

The dried precipitates were re-suspended in 50μl TE. Loading dye was added to 1× and the solution loaded onto a single 15% PAG (19:1 acrylamide:bis-acrylamide) with 10bp DNA MW markers (Life Technologies), then run at 2V/cm for 6-7hr. The gel was stained with 1×SYBR Green 1 and the bands running at 24-26bp excised and purified by eluting into 0.2M NH_4OAc and incubating at 37°C (not 65°C) for 10min (FIGURE 2.3b). The slurry was clarified and extracted as before and precipitated with 3vol of ethanol. The re-suspended tags were pooled to 20μl of 1× ligation buffer.

2.6.6 LIGATION OF DITAGS TO FORM CONCATEMERS

Purified ditags were concatenated with 0.5U/μl T4 DNA ligase at 16°C for 5hr. The entire reaction was heated to 65°C for 10min, quenched on ice then loaded onto a single lane of an 8% TAE-PAG (37.5:1 acrylamide: bis-acrylamide) and run at 7V/cm for 3-4hr. The concatemers appeared as a smear from about 100bp through to 3kb (see FIGURE 2.4). Concatemers greater then 400bp in length were excised and purified from

the gel as before (500μl TE, 65°C for 15min), then extracted, precipitated and re-suspended in 10μl of TE.

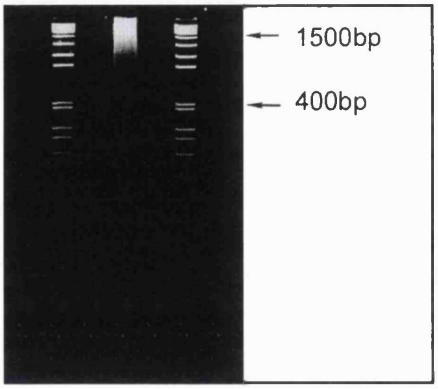


FIGURE 2.4. CONCATEMERS OF DITAGS.

Purified ditags are ligated together to form long stretches of ditags that facilitate cloning and sequencing. Following ligation the concatemers are separated on an 8% TAE PAG and appears as a smear of DNA between 100bp-3kb.

2.6.7 CLONING THE CONCATEMERS

Super coiled pZERO (1μg, Invitrogen) was made linear by digestion with S*ph*I (1U/μg, New England BioLabs) at 37°C for 30min. The digestion was extracted once with P/IAC and once with IAC then precipitated with 2vol 100% ethanol. The air-dried pellet was re-suspended in 10μl of TE (final [pZERO] 100ng/μl). Purified concatemers were mixed with the linear pZERO at a molar ratio of 1:2 and ligated with 1U of T4 DNA ligase at 16°C for 12hr. Aliquots of the ligation reaction (2μl) was used to transform 50μl of E.*coli* TOTP 10 competent cells (Invitrogen) by heat shock and were plated onto 10 low salt LB/zeocin (50μg/ml) agar plates. The plates were incubated for 14hrs at 37°C. Generally the plates would contain approximately 50-100 transformed colonies. To begin with 16 isolated colonies were analysed to determine the efficiency of ligation and transformation. Transformations were judged to be successful if >60% contained inserts larger than 400bp. All transformed colonies were re-suspended in 15μl 1×PCR buffer (see 'PCR Amplification of Vector Insert DNA') in a 96 well plate format. The samples were heated to 96°C for 10 min to inactivate nucleases, then stored at –20°C. Ligation of pZERO and concatemers were scaled up and transformed to produce in excess of 500 recombinant plasmids that contain inserts greater than 400bp. For each of three SAGE libraries from high glucose cultures and low glucose cultures this totalled over 3000 transformants.

2.6.8 PCR AMPLIFICATION OF VECTOR INSERT DNA

Transformed bacteria that had been re-suspended in 1×PCR buffer (25mM TAPS, 50mM KCl, 6.7mM $MgCl_2$, 1mM DTT, 1.0mM dNTPs) were heated to 96°C for 10min to inactivate nucleases and then cooled on ice. Inserts were amplified using a manual hot start PCR and universal M13 primers (M13for 5'-GTA AAA CGA CGG CCA GT-3'; M13rev 5'-GGA AAC AGC TAT GAC CAT G-3'). Generally 10μl PCR reactions produced sufficient products to analyse on a 1% TBE agarose gel and use in sequencing reactions. The PCR reaction was constructed by adding 2μl of the heat inactivated bacterial suspension to the base of the tube, 3μl of M13 primers (1μM each) to the side of the tube and 5μl of 2× Reddymix (0.5U T*aq* DNA polymerase, 150mM Tris-HCl pH 8.8, 40mM $(NH_4)_2SO_4$, 3.0mM $MgCl_2$, 0.02% Tween20, 0.4mM each dNTP, 2×loading dye: Advanced Biotechnologies) applied to the inside of the tube cap; the tubes were assembled on ice. When the PCR machine block had reached 96°C the reaction components were combined by centrifugation in a bench top centrifuge 1,000 rpm for 10s. The tubes were placed directly into the PCR machine and the following schedule applied.

Initial denaturation	96°C for 3min
25 cycles	96°C for 30s : 55°C for 10s : 72°C for 1min
Final extension	72°C for 10 min
HOLD	4°C

To analyse the PCR products, 2μl of the PCR reaction was loaded onto a 1.5% TBE agarose gel together with 100bp DNA ladder (Life Technologies) and run at 100V for 2hr. The gel was stained with ethidium bromide (0.5μg/ml) and visualized with UV trans-illumination (FIGURE 2.5). Products larger than 400bp were purified from excess primers and other PCR components using QIAquick columns (Qiagen) then re-suspended in 40μl 10mM Tris pH8.5. Samples were stored at –20°C until ready for sequencing.

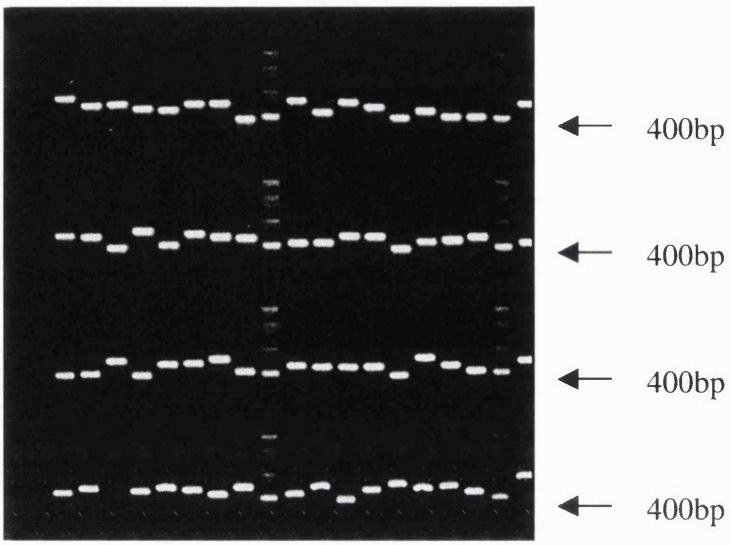


FIGURE 2.5. REPRESENTATIVE SCREEN OF TRANSFORMANTS.

Clones were chosen that contained 400bp or greater inserts. This would ensure at least 200bp of ditag cassettes. The products were purified and then sequenced. The use of the 'lethal interruption' pZERO cloning vector is reflected in the very high percentage of positive transformants and made the process more efficient.

2.6.9 DNA SEQUENCING

Double stranded DNA sequencing was conducted using the Applied Bio systems Prism DNA sequencing platform in a 96 well format. Purified PCR products (50ng) or PEG purified plasmids (1μg), were included in a 5μl BigDye terminator sequencing reaction containing 0.2μM sequencing primer (M13for 5'-GTAAACGACGGCCAGT-3'), AmpliTaq DNA polymerase FS, dNTPs and dye labelled ddNTPs. Reactions were placed in a GeneAmp 9700 PCR machine according to the following protocol. Samples were pre-denatured at 96°C for 2min followed by 25 cycles of the following schedule.

96°C 10s : 50°C 05s : 60°C 4min

A rapid thermal ramp to 4°C followed where the reactions were held until processing.

Following cycle sequencing, 15μl of 7M NH_4OAc was added to each reaction and the entire 20μl solution transferred to a 1.5ml microfuge tube. 54μl of 100% ethanol was added to the extension products and they were allowed to precipitate at R/T for 15min. The precipitated products were collected by centrifugation at 13,000 rpm for 15min. The supernatant was decanted and the pellet washed with 100μl of 70% ethanol to remove excess d/ddNTPs and primer. The extension products were dried under

vacuum in a GeneVac for 5min and, if not required immediately, were stored in this lyophilised state at -20°C.

When ready to load onto the 310 Genetic Analyser the reactions were re-suspended in 20μl of de-ionised formamide and transferred to a clean 400μl tube. Immediately prior to loading the samples were heated to 96°C for 2min then quenched on ice for 5min.

The 310 Genetic Analyser was configured for short sequencing reactions using a 47cm, 50μm i.d (internal diameter) capillary filled with POP6 polymer. Each sequencing sample was resolved in approximately 1 hr and this would provide an average of 400-500bp of quality sequence. Each output file was edited manually for software miscalling and the ASCII text file of the sequence was uploaded into the SAGE analysis software.

2.7 SAGE ANALYSIS

Text files were loaded into the SAGE 1.00 software (available from Kinsler see Velculescu et al., '95) and analysed. Each sequence file was scanned for CATG cassettes and the adjacent 10bp tags removed and compiled into a frequency list. The software is designed to take account of PCR misrepresentation by counting the same ditags only once. It has been postulated that statistically even for high abundance genes there is low probability that two tags will join together more than once (Velculescu et al., '95, Zhang et al., '97). The output from this software is a report of each tag present and its frequency of appearance. For transcriptome analysis, a report of tags and frequency for the entire experiment are created and used for mapping. The analysis was designed in three parts. First, the comparison of the three libraries in each culture series was used to determine similarity, second, the combined libraries were mapped to UniGene clusters, and finally, the two libraries were compared to one another to reveal differential tags frequencies. The final output is a table of tags and the frequency at which they are detected.

2.7.1 MAPPING TAGS AND GENES

Identification of SAGE tags was achieved using information available at the NCBI-SAGE web site (ncbi.nlm.nih.gov/SAGE). This information was downloaded and used to create a local Microsoft Access database. This database provides the information on the identity of each particular tag such as UniGene cluster ID, gene identification, mapping information, SAGE tag clustering and accession numbers for sequences held in Genebank (see TABLE 2.1). Any tag that is not present in this database, currently holding information of more than 4 million tags, is compiled for a detailed BLAST search using standard search algorithms for short fragments of DNA. There are several issues associated with the mapping data from SAGE analysis. First is the phenomenon where genes are masked by ambiguity in the tag and, second, where the tags are masked by ambiguity in the 3' region of genes. To avoid confusion these tags were excluded from further analysis. Such tags could be those that contained a high number of adenine residues, indicating close proximity to the polyA tails of transcripts, or those where transcripts from the same gene were represented by the different tags.

Dataset	Description	Contents
EST	Database of annotated EST	Single read sequences of cloned cDNA, generally 300bp
Nr	Non-redundant Database	Contains all sequences submitted. Composed of EST, STSs and characterised cDNA
TTG	Tag-To-Gene	Maps tags based on all submitted sequences
GGT	Gene-To-Tag	Extracts tags from characterised cDNA

TABLE 2.1. DATASETS AVAILABLE FOR PUBLIC ACCESS AT NCBI.

Each set has a particular use and redundancies exist. For example the EST libraries are highly redundant for high abundance genes while the nr dataset is a condensed set of representative mRNA and DNA sequences. The TTG dataset maps empirically determined tag sequences while the GGT dataset extracts tags from all sequences.

2.8 STATISTICS

The growth of expression and transcription data has resulted in a problem comparing experiments across platforms. It is reasonable to suggest that straightforward micro-array experiment may generate a million data points that cannot be efficiently sorted and analysed on a spreadsheet or graph. Similarly, SAGE generates large amounts of data but bypasses some of the technical problems associated with scanning technology and cross platform comparison in that it produces a digital

output of frequency that is directly proportional to transcript abundance. SAGE analysis is not without dilemma.

The comparison of two large datasets that were sampled from complex populations presents three issues. First is the issue of complexity. As has been reported in many SAGE libraries, profiles appear remarkably similar. It would appear that only a small fraction of transcripts represent the changes between control and experimental systems. This will result in the skewing of bulk expression correlations and mask a few significant differences by the vast majority of unchanged data points. The second issue relates to the level of data collected for each transcript. Obviously, the more data collected (in the case of SAGE the more tags seen) then the more confidence in differences seen. Again, because this generally occurs with high abundance transcripts that are constitutively expressed (e.g. housekeeping genes) then the efficiency of data collection will be lowered by the large amount of data for high abundance gene transcripts that change little. The final consideration is the determination of significance regarding the difference of tag frequency and thus abundance of gene transcript of individual tags in a complex population.

The majority of transcripts of interest will be medium to low abundance transcripts and thus library sampling level will need to be increased in order that the expression of medium to low abundance genes are sampled to a level of confidence. In the case of a SAGE library, this increase will be exponential as it is based on abundance. Increasing sampling exponentially will provide modest increases in confidence. All three of these issues are addressed in various forms in several publications, but are briefly summarised below (Chen et al., '98a, Sherlock, '00, Velculescu et al., '00, Stollberg et al., '00).

2.8.1 CORRELATION FUNCTIONS

The most popular devise to compare the similarity of two populations is a correlation analysis (Snedecor and Cochran, '73). If a data point is created from two linear metrics then these can be used to define a vector. The most common correlation used is the Pearson correlation, which is essentially a measure of the similarity between the two vectors. The Pearson correlation treats all vectors as the same unit length and so is relatively insensitive to the magnitude of the vector. The Pearson product moment

correlation ('r') is commonly used to compare signal strengths from micro-arrays or where defined subsets of large expression datasets are created. This allows a similarity to be estimated without skewing from data points falling below a magnitude threshold where differences cannot be accurately measured. A similar alternative to the Pearson product moment correlation is the Euclidean distance, which measures the difference between two points both defined by expression vectors. This was used to generate linkage graphs, which illustrate the similarity between large data sets. In this case, the magnitude and direction of the vector is incorporated into the metric. With this correlation, a large amount of data must be generated for all points of medium to low expression levels to be significant. All correlations can be used to describe similarity between individuals and relatively small populations, but are relatively insensitive to global variations.

In order for SAGE analysis to be useful, a level of confidence in both the abundance of a tag and the difference seen between the same 'tag' in two libraries must be calculated. A detailed description of the mathematical derivations can be found at the SAGEmap website and is briefly described below.

2.8.2 DETECTING A TRANSCRIPT

Assume that there are 300,000 mRNA molecules per cell and that each transcript has an equal chance of detection (Bishop et al., '74). A total of 20,000 tags are generated from each of two cell populations in a SAGE analysis. A specific species of mRNA that has a mean of N transcripts in each cell population will generate a total 'T' of {N/(300,000/40,000)=N/7.5}(Chen et al., '98a). Because each transcript of a specific type has equal chance of detection a Poisson distribution of the actual total of tags generated for each transcript can be expected. The probability that no tags are seen is e^{-T}, which is about 0.05 when T=3. Thus, a transcript that is present on average 22.5 times has a 0.95 chance of being detected. This corresponds to a relative transcript abundance of 0.056%

2.8.3 DETECTING A CHANGE IN EXPRESSION LEVEL

In two SAGE libraries a specific tag is seen with frequency A and B. In order to determine the probability that these tags differ in frequency, and so inferred mRNA

level, by a factor F, a Bayesian statistical analysis is applied where a prior distribution is used to calculate a posterior probability after a sample population is empirically determined. The Bayesian analysis has advantages over other correlative approaches in that a probability can be derived from a given factor of induction and observed frequency, but this relies on a prior assumption of probability density. In an experiment, the actual number of tags generated are Y and Z and let x = Y/(Y+Z). The proportional distribution of x is a value between 0 and 1 for all matched pairs and can be used as a prior probability density function to illustrate likelihood. For mathematical convenience, a two-parameter beta distribution is used (Kendall and Stuart, '77). The beta distribution with parameters a and b has a probability density function (pdf) described in EQUATION 2.2

$$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$$

EQUATION 2.2. BETA (*A,B*)

Where a and b are both positive and real, Γ is the gamma function and x is in the range [0,1]. A Bayesian analysis suggests that after observing A and B tags from the experiment, the posterior probability density for x is Beta (A+a,B+b). If an equal number of tags are generated in each library then for the mRNA expression to be increased by factor F, x must fall between F/(F+1) and 1 (see EQUATION 2.3). Noting that for positive integral values of n, $\Gamma(n) = (n-1)!$ the posterior probability can be calculated by integrating beta(A+a,B+b) from F/(F+1) to 1. The values of a and b are relevant for those tags for which there is little data as the prior assumptions are overwhelmed with increasing observances of A and B.

For each SAGE experiment a pair of parameters must be determined. When both a and b equal 1, the so-called 'uninformative priors', then the probability distribution becomes uniform over the interval [0,1]. This implies the probability density is equal for all pairs of tags, which is clearly not the case. Ideally parameters a and b are estimated empirically for each experiment. However, by examining the SAGE libraries produced to date, one can note that most transcripts do not significantly alter frequency between control and test libraries. This observation indicates that the probability distribution will not be uniform but will peak around x = 0.5. This suggests a and b to be equal, positive and greater than 1. The more conservative or similar the

data then the greater the probability density near $x = 0.5$ and the larger the values of a and b. Thus more information is required to infer a change in expression. Examining the total set of tag pairs will allow the estimation of a and b. More confidence in the data allows for lower values for a and b and less frequency change required to infer a change in expression. Based on current SAGE data to date prior parameters a and b have been presented to be between 2 and 4 (Lal et al., '99, Lash et al., '00).

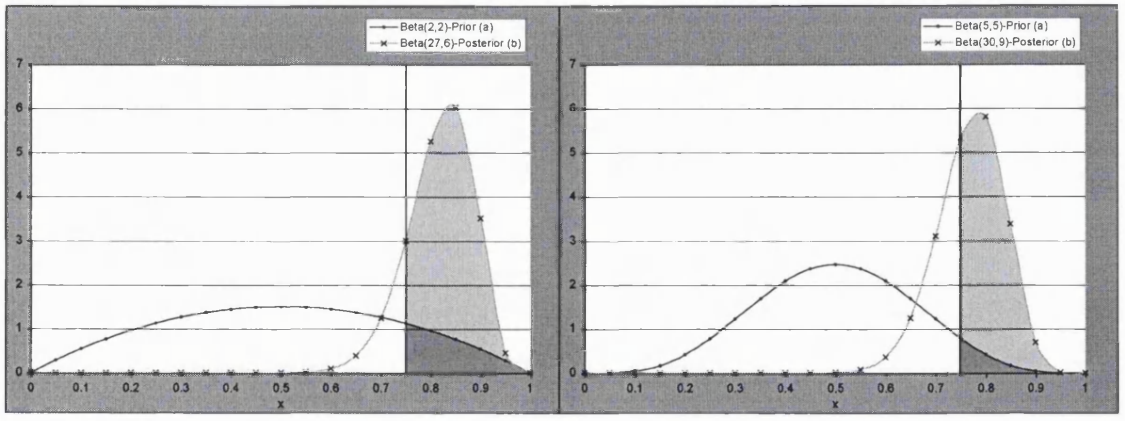
Thus to calculate the probability P that the level of a tag has been increased by a factor 3 the Beta (A+a,B+b) integral is solved between 0.75 and 1 (EQUATION 2.3)

$$P = \frac{(A+B+3)}{(A+1)!}\int_{1.0}^{0.75} x^{A+1}(1-x)^{B+1}dx$$

EQUATION 2.3. INTEGRAL OF BETA(A+*A*,B+*B*)

Illustrated graphically the prior probability density function is represented by curve 'a' (see FIGURE 2.6). The 'prior probability' that a tag is increased by a factor of 3 is represented by the area under the curve from 0.75 to 1. This is directly proportional to the integral of Beta (a,b) in the interval [0.75,1], which is 0.033 for Beta (4,4) and 0.15 for Beta (2,2). In an experiment where equal numbers of tags are generated from each of two groups, Y and Z, and a particular tag is seen 25 times in group Y and 4 times in group Z the posterior probability density function is represented by curve 'b'. The posterior probability that a tag is increased in Y by a factor of 3 from Z is the integral of the posterior PDF between 0.75 and 1, which is 0.85 using Beta (2,2) as the prior and 0.64 using Beta (5,5) as the prior. When much less data is available for a tag in Y and Z, the prior assumption becomes more important. For example, if a tag is present 10 times in Y and 2 times in Z the posterior probability that the tags in increased by at least a factor of 3 is 0.54 using Beta (2,2) as the prior and 0.26 using Beta (5,5) as the prior. The convention for using this type of analysis for determining significance is to reject tags whose probability of induction or repression by a factor of three falls below 0.5.

a



b

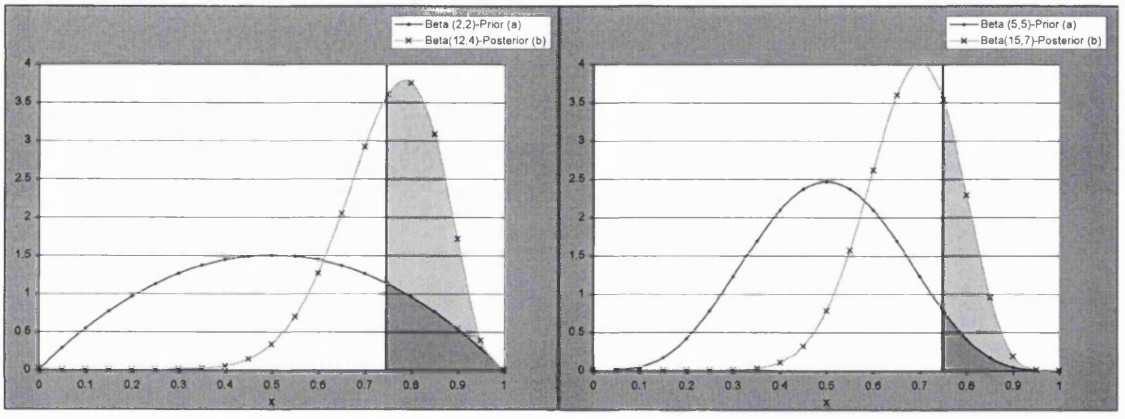


FIGURE 2.6 A & B. PRIOR AND POSTERIOR BETA PDFS.

A tag is sampled 25:4 'a' and sampled tag 10:2'b'. Integrating the pdf in the interval [0.75,1] provides the probability that the tag in altered at least three-fold. The beta parameters become less useful as more data is collected for each matched tags pairs. Using Beta(5,5) for tag pair 10:2 returns and lower probability than for Beta(2,2) Generally $p < 0.5$ excludes the tag from further analysis so as can be seen from the paired graphs the level at which a tag is sampled and assignment of prior parameters is more important than the induction level.

2.8.4 COMPARING MEANS

The largest error expected in RT-PCR experiments will be between experimental samples. Bulk mix reagents and gold standard PCR machines will minimise technical errors. Apart from cultures arising from cells obtained from an individual, all culture series were independent and semi-quantitative RT-PCR experiments were conducted in duplicate using TRNA from three independent culture series. To determine the significance between variable values obtained in real-time RT-PCR reactions the two tailed matched pair students t test was used. Significance was accepted when $p < 0.05$.

2.9 REVERSE TRANSCRIPTASE PCR

Gene specific primers were designed for each gene of interest by downloading characterised sequences from NCBI and using consensus sequences or reference sequences for primer templates. Primers were chosen based on a similar annealing temperature, to allow multiplexing of RT-PCR reactions, as well as a product of between 50-600bp biased at the 3' end of the gene and where possible incorporating the area of the SAGE tag (see APPENDIX 2).

Templates for RT-PCR were prepared by reverse transcribing 2-5μg of TRNA from the three samples at four points, proliferating cells ('P' t = -48), low glucose ('L' t = 96h), high glucose ('H' t = 96h) and equimolar mannitol ('M' t = 96h). The TRNA was diluted in DEPC treated dH_2O, then heated to 65°C for 10min before chilling on ice. The mRNA was primed by the addition of 1μg oligo-dT_{20} and allowed to sit at R/T for 2min. The primed RNA was added to a reverse transcription mix in a volume of 30 μl (50mM Tris pH 8.3, 75mM KCl, 7.5mM DTT, 10mM $MgCl_2$, 0.08mg/ml BSA, 2.4mM each dNTP, 10U/μl Murine Moloney Virus Reverse Transcriptase and 2U/μl RNAguard) and incubated at 37°C for 90min. The first strand cDNA template was diluted to 100μl and stored at –20°C.

The first strand reaction, 1μl, was included in a 20μl PCR reaction using the gene specific primers (0.5μM each) and the following cycling conditions.

Denature	96°C for 2min
35 cycles	96°C 10s : 60°C 5s :72°C 60s
Extension	72°C for 10 min

To determine the efficiency of the PCR 10μl of the PCR reaction was electrophoresed on a 1% TBE-agarose gel and examined for a single clear band of the expected size. To confirm the amplification of a gene specific product the remainder of the PCR reaction was purified from excess primers and reaction components, then DNA-sequenced using the forward and reverse primers (0.2μM). The sequence files were aligned and the consensus was used in a 'BLASTn' search to confirm the identity and size of the gene specific amplicon.

2.9.1 REAL TIME RT-PCR

Semi-quantification and relative quantification of gene expression was determined by using 'real-time RT-PCR' which monitors the generation of amplicons in real time. This is achieved by measuring the fluorescence of the dsDNA specific binding dye SYBR Green 1, which fluoresces only when bound to dsDNA and thus increases in direct proportion to the generation of amplicons in a PCR reaction. If a PCR reaction is progressing in an efficient manner (i.e. predicted product is generated in an exponential fashion), described in EQUATION 2.4.

$$X_m = X_n(1+E_X)^{m-n}$$ **EQUATION 2.4. EXPONENTIAL PCR AMPLIFICATION**

Where X_n =number of molecules at cycle n

X_m =number of molecules at cycle m

E_X =Efficiency of the amplification

m-n =number of cycles between n and m

When the efficiency is 1 (100%) then $X_m = X_n(2)^{m-n}$. So in one cycle (i.e. m-n = 1), $X_{n+1} = 2X_n$. The efficiency of the amplification reaction can be determined from the gradient of the standard curve when serial dilutions are plotted on $\log_{10}$ scale for target units versus fluorescence. A gradient of 3.322 indicates ten fold amplification every 3.322 cycles. For amplifications where efficiency is 100%, then a ΔΔCt analysis can be used to determine relative amount. If the efficiency of the amplifications is not 100% but constant then a standard curve is constructed from the serial dilutions and relative values read from this.

For example,

$10 = 1(2)^{m-n}$

$\log_2 10 = (m-n)$

$(m-n) = 3.322$

Alternatively,

$10 = 2^{(m-n)}$

$\ln 10 = (\ln 2)(m-n)$

$(m-n) = \ln 10/\ln 2$

$= 3.322$

For semi-quantification gene-specific PCR products were quantified by measuring A_{260} then diluting to a series range of 25 copies per μl (cpμ) through to 2,500,000 cpμ based on the relationship between dsDNA length, MW and molarity (1μg of 1000bp dsDNA≡ 1.52pmol). Each series was used to create a standard curve of amplification as a function of cycle number to which the experimental samples were applied. The amplification function requires the relative fluorescence captured on a digital camera as a function of cycle number determined from the point where the fluorescence crosses an empirically defined threshold value (Ct). Once assigned, this threshold value is used to determine Ct values for all samples in the experiment. Relative quantification is achieved by determining an expression level based on the expression of a ubiquitous or housekeeping gene, for example GAPDH or β-actin.

The specificity of the amplicon is determined using a melting curve analysis. The PCR product is melted by the gradual increase of temperature and the change in fluorescence measured as the strands separate. The result is a melting profile of the PCR reaction where small amplicons such as primer-dimers melt at lower temperatures than a longer specific product. This analysis serves to assess the quality of the amplification reaction and whether quantification can be considered accurate. The most specific amplifications and thus most accurate PCR reactions will have very low or no contamination with non-specific amplicons.

Two platforms of real time RT-PCR were used for the experiments described in this thesis. They were the Light Cycler (Roche) and the ABI Prism7000 (Applied Bio systems). Although the technology was essentially the same there were some subtle differences that are explained below.

2.9.1.1 LIGHT CYCLER

The thermal profiles for the Light cycler (Light Cycler Technology, Roche) were as follows,

Denature

95°C 300s (20°C/s)

Amplification

95°C 15s (20°C/s)

T_m°C 5s (15°C/s)

72°C 25s (20°C/2)

Single Acquisition (of SYBR Green1 fluorescence)

Melting Curve

95°C 0 (20°C/s)

65°C 15s (20°C/s)

95°C 0 (0.1°C/s)

Continuous Acquisition (of SYBR Green1 fluorescence)

Cooling

40°C 5 (20°C/s)

2.9.1.2 ABI 7000 SEQUENCE DETECTION SYSTEM

The technology for the ABI 7000 SDS (Applied Bio systems) is essentially the same as the Light Cycler with the following differences. The thermal profile of the PCR reaction was slightly altered with the following schedule.

Denature 95°C for 5min

40 cycles of

95°C 15s

60°C 90s

Melting curve

60°C 10s

Ramp to 96°C at the rate 1°C/min.

Continuous SYBR Green1 fluorescence detection

The efficiency of the amplification was maximised by lowering the amplicon size to under 150bp and the primers were designed using PrimerExpress software (Applied Bio systems). The use of this software allowed the multiplexing of RT-PCR reactions for several genes and an efficiency of 100% allowed the use the delta-delta Ct method for determining the relative gene expression levels based on a housekeeping and calibrator gene expression.

2.9.2 QUANTITATION OF RELATIVE OF GENE TRANSCRIPTION

The ΔΔCt method of gene expression levels compares the Ct values for all amplifications, normalises them to a housekeeping gene (ΔCt) and then compares all samples to a calibrator gene (ΔΔCt). From this analysis a value, normalised to an endogenous housekeeper and relative to a calibrator is calculated using EQUATION 2.5.

$$\text{Fold Change} = 2^{-\Delta\Delta Ct}$$

EQUATION 2.5. DELTA DELTA CT QUANTITATION

In these experiments the Ct values for all the genes in the set were normalised to GAPDH (ΔCt) and then calibrated to the 'L' sample (low glucose 96h) (ΔΔCt). In this way relative values are normalised between sample variations and expressed in terms of induction or repression from low glucose cultures. Amplicon specificity is also assessed using a melting curve analysis described above.

2.10 DIGITAL NORTHERNS

In order that the levels of tags, and thus genes, could be investigated across experimental platforms, a digital northern was constructed where a selection of SAGE libraries of related and unrelated cell types were directly compared to the NHMC library. This was used to investigate the similarity between SAGE libraries, transcriptional profiles and attempt to measure of similarity between the libraries.

SAGE libraries were downloaded from NCBI and after normalising to tag sampling levels were loaded in an Microsoft Access database. This database was termed a virtual northern as it contained transcriptional information about all the cell types loaded. The virtual northern was queried for tags that were present at high abundance in most of the libraries (presumed to be constitutively expressed genes), tags that were restricted to lineage subsets and tags that were expressed predominantly in the NHMC library. The output was termed a digital northern as it produced relative transcription information in a digital form.

A measure of the similarity between the libraries was estimated using a Pearson product moment correlation coefficient 'r'. Normalised SAGE libraries were assembled in a spreadsheet and analysed using Statistica 6 software (StatSoft).

CHAPTER 3

3 PRELIMINARY VALIDATION EXPERIMENTS

3.1 INTRODUCTION

The SAGE technique is complicated, and a substantial investment in resources. At the outset of this project, SAGE analyses were being conducted in few laboratories, so before undertaking a SAGE analysis in the NHMC model system a pilot analysis was conducted on a more robust cell system. In addition, preliminary hybridisation experiments were conducted on the NHMC protocol to assess the responsiveness of the cell system. The experiments described in this chapter had two purposes. A pilot SAGE project would confirm the technique was a viable analysis and our NHMC culture protocol was verified for RNA isolation and transcriptional response to glucose.

A small scale SAGE project was undertaken to investigate the differentiation of THP-1 cells in response to the phorbol ester PMA. THP-1 cells are an immortalised monocyte-like cell line, originally isolated from a patient with acute monocytic leukaemia. The cells are easily grown in suspension and have been used widely as a model for experiments involving monocyte activation and mitogen response. Upon stimulation with PMA, the THP-1 cells will differentiate from a transformed monocyte phenotype into macrophage like cells, indicated by the switch from cells growing in suspension, to cells that adhere to the culture flask and behave as macrophages (Tsuchiya et al., '80, Tsuchiya et al., '82). The pilot SAGE analysis would test two properties of the transcriptome. If SAGE was able to generate a valid transcriptome then the tags sampled from a SAGE library will appear in a non-uniform fashion. That is to say, abundant genes are represented by a larger amount of tags than genes of lower abundance. Secondly, genes that change abundance due to stimulation with PMA will be reflected in different frequency of tags.

To verify our NHMC model a series of hybridisation experiments were conducted to verify the transcriptional response to high glucose. These experiments had three objectives. First, to assess the quality of the RNA isolated from the culture series using Northern blotting. Second, to assess the differential transcription in the culture system, by examining previously described genes using manually constructed dot blots. Finally, medium density micro-arrays, containing some 5000 gene fragments, were used to study previously described genes and to possibly identify other genes for more rigorous analysis.

3.2 THP-1 PILOT PROJECT

3.2.1 PRELIMINARY NORTHERN BLOT

Prior to SAGE analysis RNA samples collected from proliferating THP-1 cells and PMA stimulated cells were used in a Northern blot probed with IL-1β and GAPDH. Induced transcription of IL-1β has been demonstrated in THP-1 cells after treatment with PMA, so this would serve as an indicator that differential transcription was occurring in a predictable manner (Tsuchiya et al., '82). The Northern blot demonstrated that indeed there was strong up-regulation of IL-1β mRNA and so a pilot SAGE analysis was undertaken (see FIGURE 3.1).

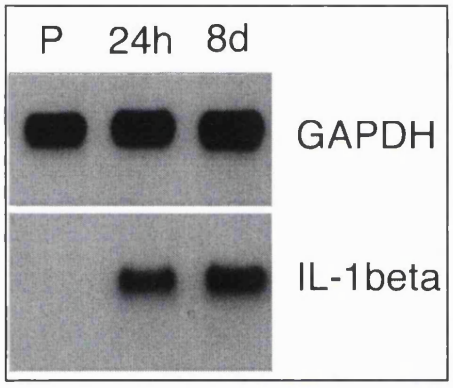


FIGURE 3.1. NORTHERN BLOT OF THP-1 RNA PROBED WITH GAPDH AND IL-1β.

Proliferating THP-1 cells (P) were persistently stimulated with PMA as described in CHAPTER 2.2.1 and RNA sampled at 24hours (24h) and 8days (8d). Persistent IL-1β induction is observed from an apparent absent level in proliferating cells.

3.2.2 TAG SAMPLING

In all, 6647 tags were sampled according to the SAGE protocol, which represent 3682 unique tags (see CHAPTER 2.6). Approximately half the tag population were derived from proliferating THP-1 cells and one half from cells stimulated with PMA for 8 days (after which the cells are considered to be fully differentiated to a macrophage like phenotype). As shown in TABLE 3.1 (a&b), the two libraries were remarkably similar in the breakdown of tag abundance, in particular the proportion of tags that were represented in abundance classes, and the proportion of tags that were present in only one library. Combining the libraries (TABLE 3.1c) revealed the bulk of the tag mass was represented by a small number of tags present at high frequency while the majority of unique tags (78.5%) were present only once.

The vast majority of tags (98.4%) showed little or no change in abundance. Of the 1.6% remaining, 54% appeared to be increased and 46% appeared decreased (TABLE 3.1d), although the discriminatory level that this was applied, i.e. greater than 3-fold induction or repression, was too low to determine statistical significance by Bayesian analysis (see CHAPTER 2.8.3).

a

THP-1	1 Tag	2 to 10 Tags	> 10 Tags	Totals	%Unique Tags in Population
Total Tags	1738	1203	578	3519	
Total Unique Tags	1738	382	27	2147	
% Unique Tags	81	17.8	1.3		
					61
Percentage of Tags Exclusive to THP-1	45				

b

THP-1/PMA	1 Tag	2 to 10 Tags	> 10 Tags	Totals	
Total Tags	1640	1222	368	3230	
Total Unique Tags	1640	363	20	2023	
% Unique Tags	81.1	17.9	1		
					62.6
Percentage of Tags Exclusive to THP-1/PMA	41				

c

Combined Libraries	1 Tag	2 to 10 Tags	> 10 Tags	Totals	
Total Tags	2888	2237	1522	6647	
Total Unique Tags	2888	721	73	3682	
% Total Tags	43.4	33.7	22.9	100	
% Unique Tags	78.4	19.6	2	100	
					55.4

d

Difference >3	Induction	Repression
Tags	32	27
% Unique Tags	0.9	0.7

TABLES 3.1(A,B,C,D). TAG FREQUENCY DISTRIBUTION FOR THE PILOT SAGE LIBRARY.

As with other SAGE libraries the vast majority of tags are present at low levels while the bulk of the tag mass is represented by a small number of tags (Table 3.1a &b). The combined libraries reduce the percentage of unique tags (a crude measure of complexity) from 62% in the individual libraries to 55% (Table 3.1c). With regard to differential frequency 32 tags were increased greater than 3-fold while 27 tags were decreased at least 3-fold (Table 3.1d)

3.2.3 MAPPING TAGS TO GENES

The 200 most abundant tags were compared to sequences in public databases (APPENDIX 3). Individual tags were aligned to the available sequences using the 'BLASTn' algorithm at NCBI (www.ncbi.nlm.nih.gov/BLAST). The filters for the 'BLASTn' alignment were set for low complexity and parameters were adjusted for

short 14bp queries. Mapping tags to genes was assigned by fulfilling the criterion of most 3' N*la* III site using only cDNA.

In all, 68% of unique tags matched entries: Of these, 73% matched characterised cDNA entries and 27% matched EST entries. The remaining 325 of tags were unmatched or derived from contaminating linker sequences. TABLE 3.2 (a&b) shows a sample of the raw mapping data derived from mapping the THP-1 SAGE library to UniGene clusters. Two levels of redundancy were observed in the SAGE library; firstly, the mapping of a single tag to multiple genes and, secondly the mapping of multiple tags to the same gene. Both sorts of redundancy confound true mapping. Of the 200 top tags, 33% mapped to more than a single gene and in the top 10 genes, 33% mapped to more than one tag. Clearly, this needs to be considered when assigning tags to genes, as it can skew true differential levels or apply false assignments. This issue will be more fully addressed in CHAPTER 5, but for the purposes of this pilot project, redundancies were removed to reveal a final transcriptome (TABLE 3.3). The proportion of genes represented in the raw data by more than one tag, or a tag that is associated with more than one gene, is summarised in TABLE 3.2b.

Removing all tag masking data reveals a transcriptome that will be a more accurate map if somewhat reduced. Examining the mapping data revealed a predictably high frequency of tags for genes associated with the cytoskeleton, such as actin, cofilin, profilin, and protein synthesis genes such as the ribosomal proteins. Other genes present at high abundance were associated with metabolism, such as GAPDH and a variety of peroxidases, as well as many transcription and translation factors (TABLE 3.3).

Interestingly, IL-1β tags were not represented in the PMA stimulated SAGE library. This may be due to the sampling population or, alternatively, the tag for IL-1β may be present but was not mapped as its frequency was too low, i.e. below n = 5. The likelihood of low level sampling or an SNP in the IL-1β tag, resulting from sequencing error or specific SNP genotype, was considered unlikely, but examined. Scanning the project for combinations of tags that contain any one SNP across the entire tag sequence returned data that suggested up-regulation of IL-1β transcription but without further analysis must still be considered inaccurate in this library. Potential tag mutations are more complex than simple SNPs. An insertion or deletion mutation can also result in expansion or contraction of the real tag sequence. Either phenomenon will result in the

potential for a wide variety of tags and large error in quantitation. Continuing the analysis with such tags would be highly complicated.

A

Tag Sequence	Total Tags	Gene Description	Redundancy		UniGene ID (Hs.)
			Genes	Tags	
ATGTCTCAAA	48	ESTs, Weakly similar to ALU7	1	1	114057
GCCTCCAAGG	37	glyceraldehyde-3-phosphate dehydrogenase	1	3	169476
AGGCAGACAG	50	eukaryotic translation elongation factor 1 alpha 1	2	6	181165
		PRO2047 protein		1	284136
CCAGAACAGA	28	ribosomal protein L30	3		111222
		KIAA0699 protein		2	17411
		deoxythymidylate kinase (thymidylate kinase)		1	79006
GATTCCGTGA	33	ribosomal protein L37	1	2	179779
CAAGGTGACA	33	ribosomal protein S2	3	7	182426
		Homo sapiens mRNA; cDNA DKFZp566D1346		1	22612
		activity-dependent neuroprotective protein		1	3657
AGAGCGAAGT	27	ESTs	1	1	246074
AAGGAAATGG	24	ESTs,	3	1	209587
		integrin, beta 1		1	287797
		Clathrin assembly lymphoid-myeloid leukemia gene		2	7885
CACAAACGGT	28	ribosomal protein S27	1	1	195453
GTGGCTCACA	20	Homo sapiens mRNA; cDNA DKFZp586J1922	6	1	138411
		Homo sapiens cDNA FLJ11095 fis, clone PLACE1005374		1	167578
		hypothetical protein FLJ20280		1	270134
		Homo sapiens cDNA FLJ14136 fis, clone MAMMA1002744		1	298014
		Homo sapiens cDNA: FLJ21554 fis, clone COL06330		1	321645
		adenosine A2b receptor		1	45743

B

SUMMARY	n
No. of Tags	10
No. of UniGene Entries	22
Genes that map to single Tags	15 (67%)
Tags that Map to single genes	5 (45%)
Unique tags mapping to single genes	3 (27%)

TABLE 3.2 A & B. TOP 10 TAGS (TOTAL MATCHES/REDUNDANCIES/MULTIPLE HITS).

Many of the high abundance genes failed to resolve discrete tag-to-gene mapping (3.2A). Only 27% of the top 10 tags mapped accurately to genes (3.2B). This preliminary finding suggested that the primary mapping of tags to genes is not accurate and will never be absolute. For example, the tag AGGCAGACAG occurs 50 times and maps to two genes ETEF1α1 and an EST. The UniGene cluster for ETEF1α1 also maps to 6 other tags in the library.

UNIQUE THP-1 MAPPING					
PMA	THP1	Tag Sequence	Total	UniGene ID (Hs.)	Gene Description
20	28	ATGTCTCAAA	48	114057	ESTs, Weakly similar to ALU7
10	27	GCCTCCAAGG	37	169476	glyceraldehyde-3-phosphate dehydrogenase
13	20	GATTCCGTGA	33	179779	ribosomal protein L37
21	9	CCCTGGGTTC	30	111334	ferritin, light polypeptide
13	15	CACAAACGGT	28	195453	ribosomal protein S27
9	18	AGAGCGAAGT	27	246074	ESTs
10	13	GCCTTTATGA	23	8768	hypothetical protein FLJ10849
7	12	AAGATCAAGA	19	1288	actin, alpha/gamma isoforms
11	8	GTGAAACTAA	19	106671	cleft lip and palate associated transmembrane protein 1
12	7	GGGTTTTTAT	19	74497	nuclease sensitive element binding protein 1
15	4	AAGGTGGAAG	19	163593	ribosomal protein L18a
9	9	ACATCATAGA	18	182979	ribosomal protein L12
10	8	AATCCTGTGG	18	178551	ribosomal protein L8
10	8	AGGACAAATA	18	172458	iduronate 2-sulfatase
11	7	CGCCGCCGGC	18	182825	ribosomal protein L35
6	11	GAATAATAAA	17	90078	nucleotide-sugar transporter
7	10	ACCAAAATCC	17	7953	HSPC041 protein
5	11	CTGAACATCT	16	75835	phosphomannomutase 1
9	7	GAAGCAGGAC	16	180370	cofilin 1 (non-muscle)
11	3	TGGTGTTGAG	14	275865	ribosomal protein S18

TABLE 3.3. TOP 20 UNIQUE TAGS (FINAL LIST AND MAPPING DATA).

Although not a complete picture of the transcriptome, this table lists the accurate mapping of tags to genes and is thus considered the top 20 of the final transcriptome. While other tags may represent genes, each tag only maps to one gene.

3.2.4 DIFFERENTIAL GENE EXPRESSION

A total of 59 tags, (1.6%) exhibited greater than three-fold difference in frequency. The top 40 differentially expressed tags are shown in TABLE 3.4. Very few tags reached induction or repression levels that were required to confer statistical significance and even increasing confidence in the data, thus flexibility in the statistical analysis, an induction or repression of 10-fold would be required for significance. This occurred in only 2 tags, one of which mapped to more than 5 genes the other to an adenylate cyclase associated protein that was represented by more than 50 tags (see TABLE 3.4). Due to the subjective nature of the statistical analysis (see CHAPTER 2.8.3), other tags that were induced and repressed will be present, but require further sampling in order to confirm this. It was considered unhelpful to attempt to resolve these redundancies further in the context of this pilot project.

3.2.5 CONCLUDING REMARKS ON THE PILOT PROJECT

Two SAGE libraries were constructed that represented the transcriptional profile of proliferating THP-1 cells and the same stimulated by PMA. The data was informative in that high abundance genes were more likely to be sampled than low abundance genes irrespective of whether they are affected by PMA, e.g. metabolic enzymes such as GAPDH and cytoskeleton genes, such as actin. Because of the pilot project, SAGE was considered technically feasible for use in a larger project. However, potential analyses problems were identified, such as the redundancies in the accurate mapping of tags to genes.

PMA	THP1	Tag Sequence	Ind/Rep	UniGene ID	'p' (Beta (4,4))	Gene Description
10	0	AAGGTAGCAG	>10	104125	0.6	adenylyl cyclase-associated protein
9	0	TTGGGGTTTC	>9	62954	0.6	ferritin, heavy polypeptide 1
8	1	TGGCCCCAGG	8	268571	0.4	apolipoprotein C-I
8	0	CTAAGACTTC	>8		0.5	No Match
6	1	ATTTAGAGGT	6		0.3	No Match
6	1	GCCGTGTCCG	6	241507	0.3	ribosomal protein S6
6	1	TGGGCAAAGC	6	2186	0.3	eukaryotic translation elongation factor 1 gamma
10	2	GGATGCTGGG	5	8438	0.3	ESTs
5	1	GCGACGAGGC	5	2017	0.2	ribosomal protein L38
5	1	TGGCGTACGG	5	ribo	0.2	Tag matches ribosomal RNA sequence
5	0	GTGAAGGCAG	>5	77039	0.4	ribosomal protein S3A
15	4	AAGGTGGAAG	4	163593	0.3	ribosomal protein L18a
16	4	TGTGTTGAGA	4	181165	0.4	eukaryotic translation elongation factor 1 alpha 1
11	3	TGGTGTTGAG	4	275865	0.3	ribosomal protein S18
8	2	TAAGATCCTT	4		0.2	No Match
4	1	AACGAGGAAT	4		0.2	No Match
4	1	AGCAATTCAA	4		0.2	No Match
4	1	CCTACTAACC	4	128873	0.2	ESTs, Highly similar human Fructose bisphosphate aldolase A
4	1	CTCTTCGAGA	4	76686	0.2	glutathione peroxidase 1
1	6	GTAAGCATAA	-6		0.3	No Match
0	6	GTAAGCAAAA	<-6		0.4	No Match
3	15	AACAATTTGG	-5		0.4	No Match
1	5	AACGCTGCCA	-5	301011	0.2	KIAA0876 protein
1	5	ATCCGAAAGA	-5	322804	0.2	EST
1	5	TCTGGACGCG	-5		0.2	No Match
0	5	AAGGACATCA	<-5	98110	0.4	ESTs
2	8	AAAACAGTGG	-4	5566	0.2	ribosomal protein L37a
2	8	TGAGCAAAAG	-4		0.2	No Match
2	8	TGGGTTGTCT	-4		0.2	No Match
1	4	ATGTGGTGTG	-4	180909	0.2	peroxiredoxin 1
1	4	CCAGTCCTGG	-4	221166	0.2	ESTs
1	4	GAAATATATG	-4	429	0.2	ATP synthase,subunit c isoform 3
1	4	GACTGAATCT	-4	37936	0.2	suppressor of variegation 3 homolog 1
1	4	GCGAAGCTCA	-4		0.2	No Match
1	4	TGGCTCGGTC	-4		0.2	No Match
1	4	TTGGCTGCCC	-4		0.2	No Match
1	4	TTGTGCAAAA	-4		0.2	No Match
0	4	ACATACAATT	<-4		0.3	No Match
0	4	CACAGACTGT	<-4	143863	0.3	Homo sapiens chromosome 18 unknown mRNA sequence

TABLE 3.4. TOP 40 DIFFERENTIALLY EXPRESSED TAGS.

High abundance genes are more likely to have several tags associated with them and so the presence of high abundance genes together with a low p (Beta(4,4)) values suggested that this was not an accurate reflection of differential transcription. Fold induction (Ind) or repression (Rep) are indicated as +'ve and –'ve values respectively.

3.3 NORTHERN BLOT HYBRIDISATIONS

Northern blots were used to examine the quality of RNA isolated from NHMCs cultured in quiescent conditions and under glucose stress, and to verify that transcription was occurring in a predictable fashion. Three genes were used to probe Northern blots,

selected based on reports of differential transcription following glucose stress. The facilitative glucose transporter, GLUT-1, has been previously been described to be up regulated in response to glucose in rodent cells, and within the time scale used in this culture model (Heilig et al., '97a). Thrombospondin 1, THBS1, has been previously described increased in normal human mesangial cells although outside the time scale of this culture series (Murphy et al., '99). Finally, glyceraldehyde-3-phosphate dehydrogenase, GAPDH, was used as a high abundance gene to normalise the well-to-well variations in RNA loading. Probes were selected from UniGene clusters and constructed as random labelled PCR products from IMAGE EST clones.

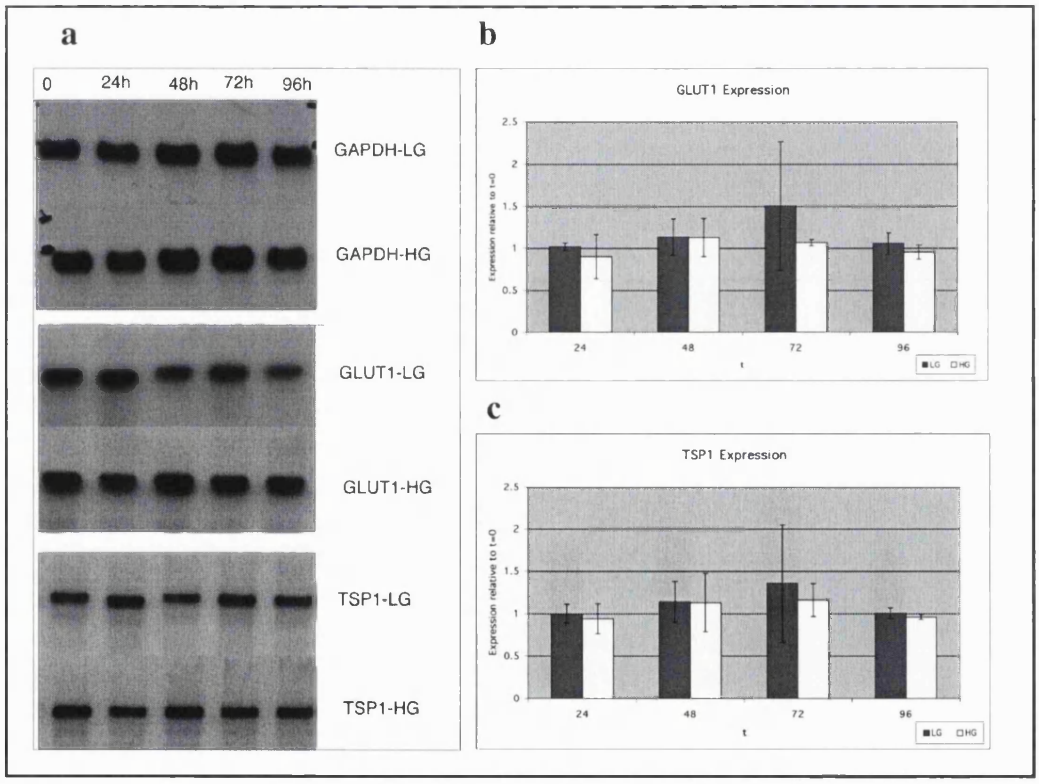


FIGURE 3.2A, B & C. NORTHERN BLOT OF THBS1 & GLUT1.

5μg of total RNA per well isolated from NHMC grown under normal (LG) or hyperglycaemic (HG) conditions for 4 days (a). Probes were derived from clones (Accession No.), M17851 (GAPDH), AA044067 (GLUT1) and AI078816 (THBS1). Tight banding patterns indicate that the RNA was not degraded and densinometric scanning from three hybridisation experiments revel no significant change between gene expression of TSP1 (b) or GLUT1 (c).

Signals from hybridised northern blots displayed tight banding patterns consistent with un-degraded RNA (see FIGURE 3.2). This was present across the entire time course. When normalised to the housekeeping gene GAPDH, the transcriptional abundance of GLUT1 and THBS1 did not alter significantly. This was of interest because GLUT1 in particular was expected to show at least a twofold increase in abundance within the first 48hrs.

3.4 DOT BLOT HYBRIDISATIONS

A further selection of genes, previously described to alter abundance because of glucose stress, was obtained through UniGene IMAGE annotations from HGMP. PCR amplified inserts for the genes were fixed to nylon membranes at various dilutions and hybridised to complex cDNA probes generated from mRNA isolated from proliferating cells and cells grown for 96h in low glucose (L), high glucose (H) or mannitol (M). Density scanning was used to compare signals from the same targets on each of three hybridisation filters. Each gene was represented by up to six separate points and thus a more accurate measure of error could be determined.

The dot blot data appeared to concur with the predicted pattern of gene expression. Although not quantitative in the setting of hybridisation of complex probes to simple targets, high abundance genes such as actin and fibronectin generally provided stronger signals than medium to low abundance genes such as specific collagen isoforms and TGFβ1. Four genes were used to verify that differential transcription was occurring, smooth muscle actin (α2-SMA), the classic indicator of DN transforming growth factor beta1 (TGFβ1), type IV collagen α2 isoform (COLLIVα2) and fibronectin 1(FN-1), ECM components whose expression in MC is increased in high glucose environments and in DN. Expression of these genes were increased in this culture system, but not to the degree reported elsewhere (see FIGURE 3.3).

The transcription of all three genes was increased approximately 40% between low glucose and high glucose cultures ($p<0.05$ TGFβ1, $p<0.1$, FN1 and COLLIVα2). An even greater difference was seen between proliferating cells and cells grown under high glucose, presumably due to removal of mitogens or matrix factors present in the FBS containing culture medium. Using a selection of housekeeping genes to normalise the experimental variation, the data generally produced similar results and so this was considered sufficient for the purposes of this project. It is generally accepted that dot blot data of low resolution can yield inconsistent data regarding measuring differences, and is difficult to optimise for detection of less than several fold induction or repression. Nonetheless, significance was determined from the repetition of experiments (FIGURE 3.3).

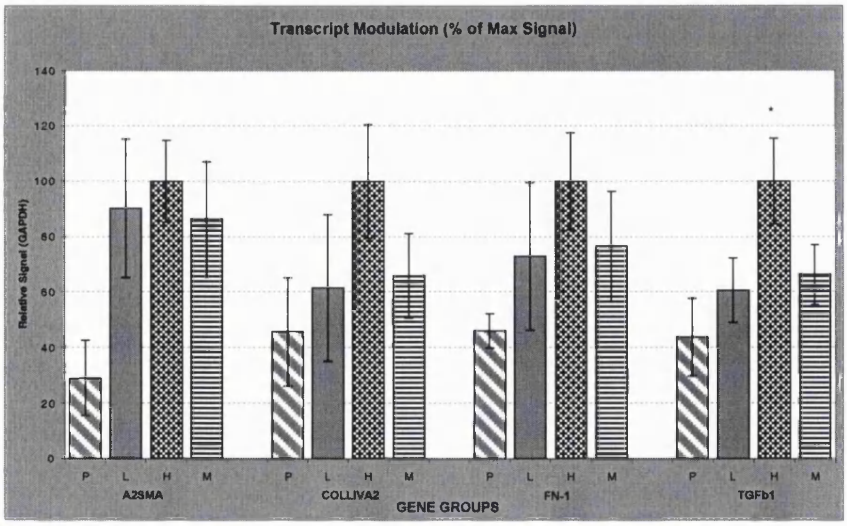


FIGURE 3.3. MODULATION OF SELECTED GENES BASED ON DOT BLOT DATA.

Targets (αSMA, COLL IVα2, FN1 and TGFβ1) were hybridised to labelled RNA from proliferating cells (P), cultures grown for 96h in low glucose (L), high glucose (H), and equimolar mannitol (M). Because of large variation in signal intensities values are expressed as relative to GAPDH and a percentage of the strongest signal in each group. A general increase of 40% in transcription of TGFβ1 was observed which concurs with previous reports in this time scale (* L v H $p<0.05$). Errors are expressed as a standard deviation from the mean of 9 signal points from three experiments.

3.5 GENE FILTER HYBRIDISATIONS

Micro-array analysis provides a further level of resolution and using this approach we were able to monitor 5000 genes simultaneously. Hybridisation protocols are similar to standard techniques and the data extracted should be able to predict a modulation of at least two-fold. A complete list of the gene fragments for GeneFilter 200, used in these experiments, is available from the manufacturer (Invitrogen/Research Genetics) but the top 200 genes detected in the GeneFilter analysis is presented in APPENDIX 6. A subset of genes was actively monitored for the purposes of culture validation in three hybridisation experiments, using RNA isolated from the same time points but in different experimental series (see FIGURE 3.4).

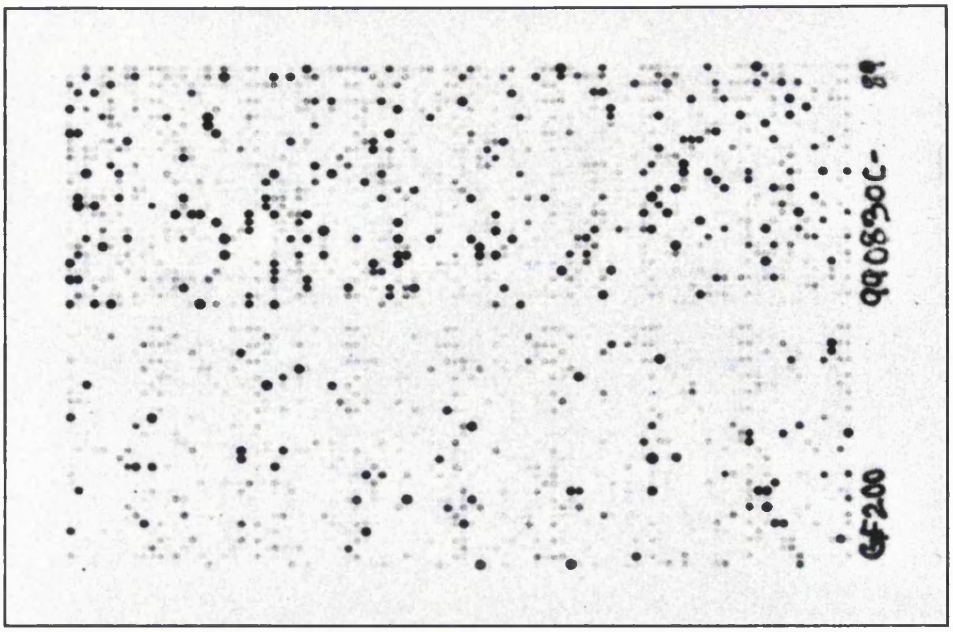


FIGURE 3.4. GENE FILTER GF200 HYBRIDISED TO 2μG OF LABELLED FIRST STRAND CDNA.

The signals are quantified at maximum resolution using a Storm 860 Phosphor Imager (Molecular Dynamics). Pathways software (Research Genetics) tracks point intensities, then compares corresponding points on other experimental filters. This particular filter was hybridised to ^{33}P labelled cDNA reverse transcribed from 2μg of total RNA. The filter was exposed to a low energy phosphor screen for 72hr. This image is typical of all the images captured and shows discrete points signals of varying levels across the entire membrane with no blurring, smudging or background. This indicates a robust level of experimental technique.

Differential transcription was observed, although again not to the degree expected and of low significance. A 'greater-than-2' convention is applied to this type of data were magnitude fold change below ± 2 is regarded insignificant. Differential transcription of genes not previously reported by glucose stress was also observed. Because of the nature of this 'bottom up' approach to data acquisition, there is much more data available for genes other than those reported in similar culture models. These data are represented on a scatter graph (see FIGURE 3.6) and compiled for further analysis based on their persistent difference in hybridisation signal intensity (TABLE 3.6). For the purposes of this thesis, these will be addressed in CHAPTER 7.

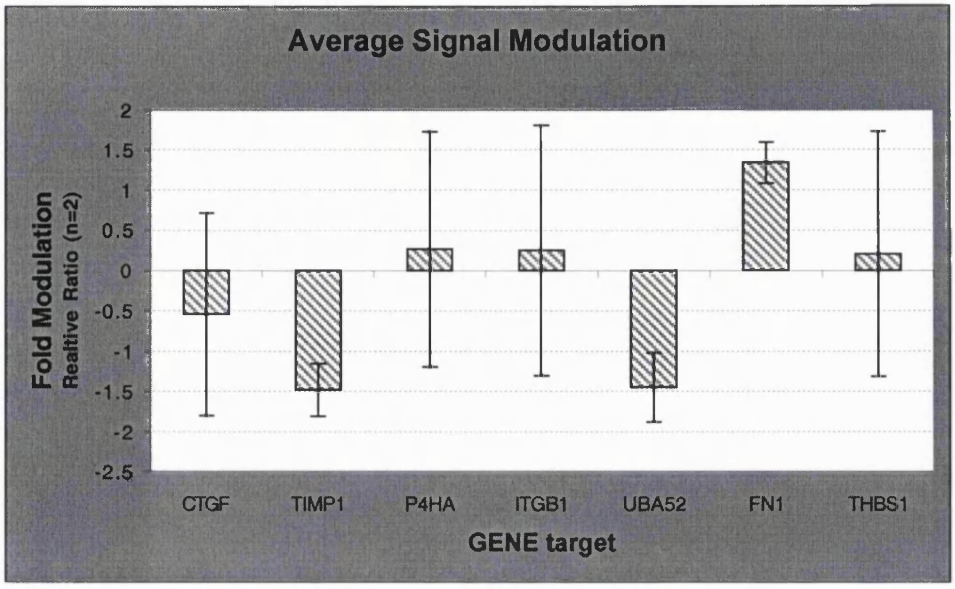


FIGURE 3.5. MODULATION OF SELECTED GENES RELATIVE TO LG ON GF200.

A selection of genes that were actively tracked and shown previously to be modulated in mesangial cells grown under high glucose. These were connective tissue growth factor (CTGF), tissue inhibitor of metalloproteinase 1 (TIMP1), proline-4-hydroxylase (P4HA), Integrin β1 (ITGB1), Fibronectin 1 (UBA52 & FN1) and thrombopondin1 (THBS1). Error bars are represented as the standard deviation from the mean of three experiments. A general suppression of transcription in high glucose is seen which is contrary to modulation based on previous reports and could possibly indicate an error in normalisation. Signal intensities for ITGB1, TIMP1 and THBS1 were considered borderline for Gene Filter resolution.

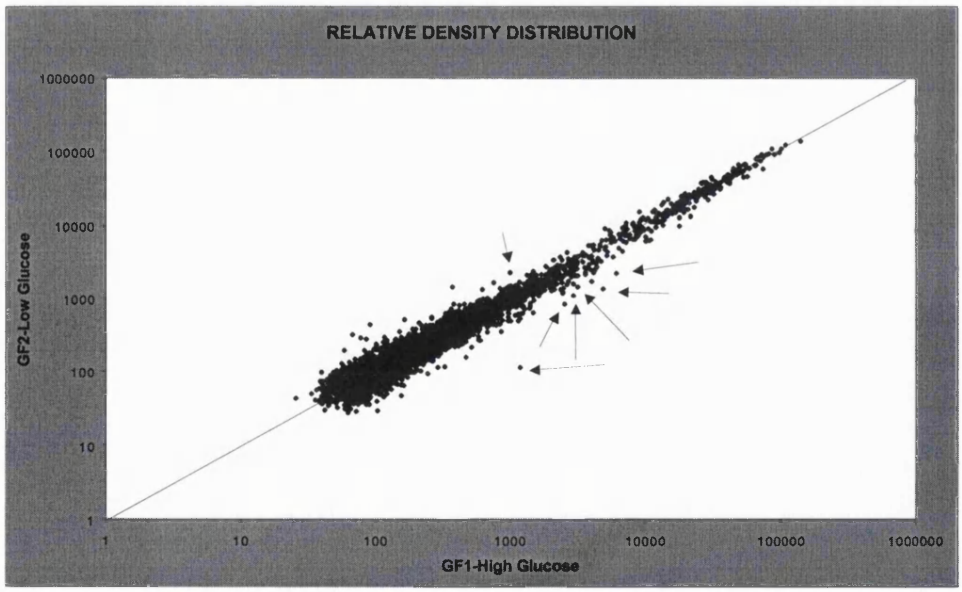


FIGURE 3.6. DENSITY DISTRIBUTION OF RELATIVE SIGNALS FROM GF200.

All data-points from two Gene Filter experiments from low and high glucose cultures. The tight clustering of data points around the 'line of similarity', whose gradient equals 1, suggests that the two culture systems are very similar. Points indicated by arrows were a selection of persistently altered in two hybridisation experiments, see TABLE 3.6, but for the majority of data there is a remarkable similarity between the two experimental conditions. Relative coordinate signals above 1000 were considered reliable.

As with the dot blot experiments the genes of high abundance, housekeeping genes, cytoskeleton genes and transcription factors, generally produced stronger signals than genes of medium to lower abundance. Although the quality of the RNA was assessed in the previous section, it was possible that degradation or inefficient labelling of cDNA occurred. However, if this were the case one would expect similar point signal strength of low intensity across the entire membrane. This was not seen, and as such, it was concluded that indeed the signals were valid.

Gene	UniGene ID	Accession	Ratio 1	Ratio 2
Fibrinogen Gamma-A Chain	Hs.75431	T94626	-1.04	2.04
FGFrc3	Hs.1420	AA417654	1.02	2.21
CD58	Hs.75626	AA136359	2.92	3.57
KIAA0146	Hs.74670	AA401448	1.91	2.02
APPBP2	Hs.84084	AA046411	-1.98	-2.20
Seb4D	Hs.104642	AA459588	1.12	2.30
HREV107-like	Hs.37189	AA476543	1.01	2.99
ESTs	Hs.101490	H51056	-2.41	2.73
ESTs	Hs.55452	W31784	1.46	2.31
ESTs (Hs.233634)	~~Hs.48217~~	W77990	2.09	2.33
ESTs	Hs.42029	H94936	-2.17	-3.96
ESTs	Hs.20261	R27432	1.05	-2.15
ESTs (Hs.326725)	~~Hs.23596~~	R23924	2.03	2.64
ESTs (Hs.183655)	~~Hs.12166~~	W47156	1.94	2.18
ESTs	Hs.12097	R99311	4.11	10.26

TABLE 3.6. GENES DIFFERENTIAL REGULATED AS ASSESSED BY GENEFILTER ANALYSIS.

As with convention a persistent average 'greater-than-2' fold induction or repression is used to filter the raw data and are indicated here in boldface. Note also the incorrect assignments of UniGene cluster ID that indicates the dynamic nature of the UniGene clustering process and as such accession numbers for the specific clones used in the Gene Filter and any subsequent analyses.

3.6 DISCUSSION

Described in this chapter were preliminary experiments designed to test the feasibility of the SAGE technology, assess the robustness of the culture protocol and test the reproducibility of the transcriptional response to high glucose. Because the SAGE technique requires significant investment in time and resources, and because at the onset of this project only a few laboratories worldwide were using SAGE, it was considered important to confirm the ability to conduct SAGE. This was achieved by assessing the sampling a pilot SAGE library and analysing the frequency distribution of tags and mapping the identity of tags to genes. Because the analyses would generate large amounts of high-resolution data, and the literature describes many culture protocols using mesangial cells and glucose stress, it was considered important to test

the robustness and reproducibility of differential transcription using the culture protocol described in CHAPTER 2. The number of genes that can be simultaneously analysed is different for a particular hybridisation technique, and is essentially dependent on the support for targets and the imaging technology available. Northern blots can accurately analyse several genes at once but the technique is cumbersome and requires relatively large quantities of RNA. Dot blots are less sensitive to transcript variation and background noise but can simultaneously analyse 2-5 genes per cm^2 and can be imaged with standard autoradiography. Medium to high-density gene arrays can analyse from 200 to 10,000 targets per cm^2, but access to the infrastructure to process data is currently expensive. The actual number of targets that can be accurately quantified is now dependent upon the resolution and sensitivity of the imaging technology

The pilot SAGE project demonstrated that the SAGE procedure was robust and produced data of similar quality to other SAGE libraries with regard to the frequency distribution of tags. This demonstrated that tags were not sampled with a uniform probability and the frequency distribution they created was similar to previous SAGE libraries. The THP-1 library contained predictably high levels of tags that represented high abundance genes, such as genes associated with the cytoskeleton, metabolism and protein translation.

The data proved less informative with regard to differential transcription of genes. Of particular note was the absence of the tag for the gene IL-1β. It was predicted that this tag would be present in the stimulated population, but absent in the non-stimulated population. As it happened, the tag that maps to IL-1β was not found in either library. This was most likely due to the low level of sampling, reflected in the statistical analysis. A recent study conducted a similar experiment using normal monocytes stimulated with LPS. A SAGE analysis was performed following 3hr stimulation with LPS, and from the 35,700 tags sampled, an increase in IL-1β of some 20-fold (9 v 177) was observed (Suzuki et al., '00). This corresponds to a presence of about 14 tags when normalised to the level of this pilot study. While it is problematic to compare transcription between normal and transformed cells, it is possible that IL-1β may not have been sampled in this THP-1 library and thus represents an error in tag generation, or is simply beyond the resolution of sampling in this scale using these culture conditions. Furthermore, the THP-1 library represents a persistent stimulation with PMA, rather than the sudden stimulation described in the study by Suzuki et al

2000. It is possible that levels of IL-1β gradually decrease with persistent PMA stimulation over the time scale used in this experiment, but this is not apparent from the Northern blot (FIGURE 3.1).

The data generated from the three independent techniques used to validate the NHMC culture system and RNA isolation proved unclear. Northern blots failed to show differential transcription of THBS1 and GLUT1. An alternative dot blot technique demonstrated differential transcription in the classic indicator of DN, TGFβ1, but the GeneFilter experiments, while identifying alternative genes for analysis, failed to demonstrate constant altered transcription of candidate genes present on the filter.

Northern blotting was primarily used to test the quality of the RNA isolated. The genes used to probe RNA from this culture have previously been shown differentially transcribed in similar culture systems. Their failure to show differential transcription in these experiments may be due to differences in time scales and conditions compared to previously published data (Murphy et al., '99). Longer time scales were not used in our model, based on advice from the distributor (BioWhittaker UK). First, the NHMC cells were used at low passage as they begin to show suppressed growth after passage 8, and second, periods longer than four days in serum free media were not recommended.

The dot blot data appeared to mirror previously reported modulation of TGFβ1. All genes tested, with the exception of α2-smooth muscle actin, have previously been reported to increase transcription in response to glucose but only TGFβ1 was statistically significant. The other genes suggested a trend of increasing transcription. The similarity displayed between the individual expression profiles is a result of the method of comparison used. Individual genes are not transcribed at the same level and so their signals were of very different magnitude. It was for this reason of magnitude that the graphing data was calibrated to the strongest signal in each set rather than relative to all signals.

The GeneFilter analysis provided the highest resolution with regard to the number of genes that could simultaneously be monitored. Despite this resolution, not all the known candidate genes modulated in glucose stress are present on the filter. Of the genes that were present, the significance of the difference in hybridisation signal

was questionable. The induction or suppression of transcription for all the actively monitored genes was below two-fold and so considered insignificant (as instructed in the GeneFilter user manual). Yet, a differential pattern of transcription was apparent for other genes. This indicates several possibilities. First, that the GeneFilter data is invalid as neither the actively tracked genes nor the differentially transcribed genes could be resolved. Second, that the GeneFilter data is real and that the actively tracked genes are not responding in a predicted manner, but the observed data on the genes is valid. Third, that the GeneFilter data is experimentally unstable and requires additional experiments to obtain accurate data. Certainly more data was expected from the 5000 genes present on the filter but because of the enormous differences in signal intensities, not all of the genes could be tracked. Approximately 1/5 of the genes present on the filter produced a signal that was within the threshold of resolution as ascribed by the manufacturer.

Taken together the three techniques of validation; northern blot, dot blot and GeneFilter indicate that while the predicted transcription pattern is not occurring to the degree expected there appeared significant differential transcription. Despite the inability to resolve most of the previously described genes, differential transcription was noticed in genes that had not formed part of the actively monitored subset. This may indicate two situations. First, the possibility that despite culturing primary MC of low passage under high glucose there is no modulation of expression for these genes within the time scale of 96 hours. Certainly there is variety of culturing protocols used for *in vitro* models in the literature, some of which extend beyond 4 days (sometimes as high as 2 months) under high glucose. Second, the observed modulation in transcription occurs prior to the altered transcription classically described in MC cultured under high glucose. Both these situations are possible, though the large body of data indicating differential transcription of the selected genes within this *in vitro* model suggests that the latter may best describe this particular model.

Taking all the results together there is clearly variation in the data. It seems unlikely that experimental variation is responsible for these errors as all the techniques suggest that the two culture systems are remarkably similar. It would appear that either gene transcription varies little in this model, or un-described genes await characterisation. SAGE analysis offers more data than simply differential analysis and so both these questions may be addressed.

A SAGE analysis will assist in clarifying both the issue of altered candidate gene transcription and identification of novel gene transcription. The nature of SAGE, based on transcript abundance rather than pre-determined sequence, not only permits the active monitoring of candidate genes but also allows the identification of novel genes involved in the model system. In addition, SAGE offers more than simply a technique to identify differentially transcribed genes. A catalogue of genes will be generated based on the abundance of their transcripts and so will constitute a transcriptional profile of the NHMC that can be used as a resource for future study.

CHAPTER 4

4 CONSTRUCTION & SAMPLING OF NHMC SAGE LIBRARIES

4.1 INTRODUCTION

Experiments in CHAPTER 3 demonstrated that constructing SAGE libraries was technically feasible and data of similar quality to published SAGE libraries was generated. Tags were generated from mRNA and were randomly sampled but displayed a non-uniform distribution. High abundance genes, such as housekeeping genes, were represented by high frequency tags, and the majority of genes were of low abundance thus the SAGE analysis appeared to reflect the transcriptome, and was considered valid. In addition, transcription in the culture protocol was shown to respond to glucose stress and a set of genes was identified for further analysis. Based on these results a large scale SAGE project was undertaken to generate a high-resolution transcriptional profile of NHMCs grown under low and high concentrations of glucose.

Establishment of the NHMC cell culture was described in CHAPTER 2.2.2 and tested in CHAPTER 3. Three SAGE sub-libraries were generated from independent cultures grown under either normal or high glucose. Each of these six sub-libraries were analysed for similarity and then each set of three sub-libraries were combined to form the primary NHMC SAGE libraries. Reasons for sampling three independent libraries for each experimental condition were two-fold. Firstly, sampling independent cultures would provide an opportunity to indicate consistency between the cultures. Any serious variation in culturing would be observed as non-linear accumulation of tags in one sub-library compared to the next. Secondly, this stability of tag generation would assist in filtering candidates for differential screening, complementing the application of Beta analysis (and described later in CHAPTER 7). The nature of sampling complex populations facilitates significant variation in the sampling of low abundance tags within each sub-library. Thus, stable differential frequencies within each sub-library, and between each experimental group, offers a more likely 'real' frequency of tags and thus more accurate estimation of the differential transcription of genes.

Each sub-library was sampled using 500 transformants and each of these sets was analysed in the SAGE software to produce independent sub-library tags lists. Six sub-libraries produced 46,219 tags with approximately half from low glucose and half from high glucose culture conditions. It was postulated that SAGE libraries would represent the transcriptional profile of cultured NHMCs, and may be used as a basis for classification.

4.2 CUMULATIVE SAMPLING OF TAGS

To monitor the complexity of the sampling process the accumulation of all tags was tracked through the sequencing project and remained approximately linear to the number of tags sampled, as illustrated in FIGURE 4.1. New tags were being sampled at a constant rate even as the sampling population exceeded 45,000 tags.

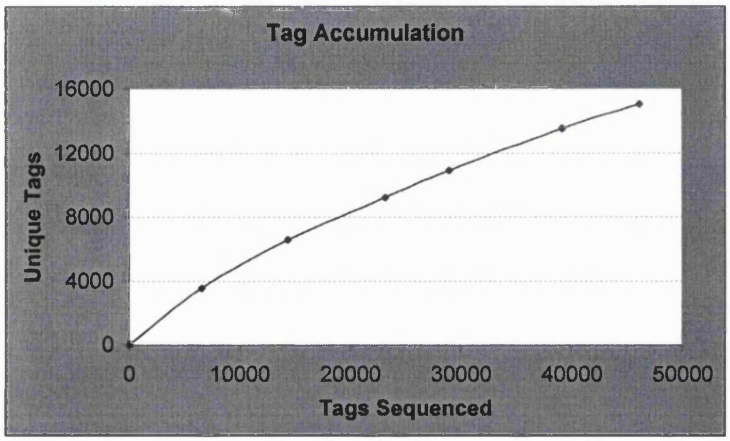


FIGURE 4.1. THE ACCUMULATION OF UNIQUE TAGS AS A FUNCTION OF TAGS SAMPLED.

From the graph it appears that sampling of unique tags is linear to the number of tags sampled. An inflection in the curve can be noticed from 0-15,000 tags yet constant novel sampling was still occurring beyond 45,000 tags.

A summary of accumulated data from the individual sub-libraries is presented in TABLE 4.1. All sub-libraries appeared similar with regard to the percentage of contaminating linkers, but there appeared small differences in efficiency, as measured by the percentage of duplicates and tags generated in each clone, (e.g. LG1 v LG3). Duplicate ditags are considered PCR artefacts and potential bias in a SAGE analysis. The likelihood that two tags will join in the protocol is considered so rare as to only count the same ditags once in the analysis (Velculescu et al., '95, Zhang et al., '97).

NHMC Sub-Libraries						
Culture set	Clones Seq	Tags Sampled (Unique)	Linkers (%)	Total Tags Linkers Removed	Duplicates	Tags per clone
HG1	452	6525(3559)	179 (2.7%)	6346(3545)	759(23%)	14.4
HG2	427	7813(4233)	214 (2.7%)	7599(4219)	766 (19%)	18.3
HG3	455	8790(4550)	241 (2.7%)	8549(4533)	1036 (23%)	19.3
LG1	451	5838(3189)	176 (3.0%)	5662(3174)	448 (15%)	12.9
LG2	453	10294(5095)	280 (2.7%)	9924(5072)	576 (11%)	22.5
LG3	305	6959(3736)	164 (2.3%)	6795(3724)	471 (13%)	23

TABLE 4.1. GENERAL EFFICIENCY STATISTICS OF THE INDIVIDUAL SAGE LIBRARIES.

Tag generation appeared to be generally stable across the respective libraries (average of 7703 in each sub-library). Variation in efficiency and contamination is reflected in 'Tags per clone' values and the frequency of linker sequences present (an average of 2.7% linker contamination was measured).

4.2.1 TAG SAMPLING INDICATES A COMPLEX POPULATION

As the sub-libraries were sampled, the distribution of tag frequency was analysed with a simple step distribution (TABLE 4.2 a & b). Studies on mRNA abundance have described discrete abundance classes within the transcriptome (Bishop et al., '74). Similar abundance classes were observed in original SAGE studies and many further studies have adopted classes based on original SAGE analyses (Velculescu et al., '95, Velculescu et al., '97, Zhang et al., '97). In the complex system of gene transcription, it is likely that the real frequency distribution of mRNA transcripts will be specific for each cell type and state, but in any case will not be uniform. The best calculations will always be mathematically derived from discrete functions but remain essentially estimates.

a. Low Glucose

Distribution	LG1 Freq	LG2 Freq	LG3 Freq
Absent	5860 (65%)	3962 (43%)	5310 (59%)
1	2546 (80%)	3966 (78%)	2961 (79%)
2 to 5	513 (16%)	886 (18%)	607 (16%)
6 to 15	78 (2%)	151 (3%)	111 (3%)
> 15	37 (1%)	69 (2%)	45 (1%)

b. High Glucose

Distribution	HG1 Freq	HG2 Freq	HG3 Freq
Absent	5667 (62%)	4993 (54%)	4679 (51%)
1	2805 (79%)	3338 (79%)	3518 (77%)
2 to 5	594 (17%)	727 (17%)	823 (18%)
6 to 15	110 (3%)	109 (2%)	138 (3%)
>15	36 (1%)	45 (1%)	54 (1%)

TABLE 4.2 A & B. THE FREQUENCY DISTRIBUTION OF TAGS IN EACH SUB-LIBRARY.

The distribution classes have been described previously and are based on empirical data (see text). Each of the sub-libraries displays remarkably similar levels of tags for each of the abundance classes. Absent tags are based on the total population of tags for the combined libraries (LG-9034 tags and HG- 9212 tags) and indicate the degree of complexity of the population.

Upon clustering the tags into abundance classes, there was a high level of similarity between the various sub-libraries and the two experimental systems (low and high glucose). The similarity between sub-libraries and experimental conditions may indicate two things, first that the three sub-libraries demonstrate little variation between culture series, thus indicating robust experimental conditions, and secondly that the

similarity between the experimental systems suggests little effect of glucose on the global transcription dynamics in NMHCs. Both these observations are unsurprising as it is probable that only a small number of genes will change abundance in response to stimuli, and these will not be of significant magnitude to affect the overall transcriptome. The similarity between control and experimental SAGE libraries has been noted in many SAGE analyses, including activation of mast cells and monocytes (Chen et al., '98a, Hashimoto et al., '99), infecting cells with viruses and transformation vectors (Polyak et al., '97, Kenzelmann and Muhlemann, '00), and between normal and cancerous tissue (Zhang et al., '97, Hibi et al., '98).

4.2.2 EACH SUB-LIBRARY HAS SIMILAR COMPLEXITY

To test whether there were similar levels of tags in each sub-library, a database was created that scanned each of the three sub-libraries for the same tags. To illustrate the complexity of the libraries Venn Diagrammes were constructed (see FIGURE 4.2). Correlation coefficients were calculated from the levels of each tag present in each set of two sub-libraries (intersection of each sub-library). The weight of the majority of tags that are present at low frequency (e.g. 0,0,1) may skew any correlation calculation and so only tags for which there was a large amount of data were used for these correlation analyses. While this is not complete or accurate, the presence of high levels of individual tags in only one sub-library was considered rare.

When the analysis was repeated using data from the intersection of all sub-libraries the results differed little. This would be expected as the intersection will favour the use of constitutively expressed, or housekeeping genes present at about the same levels in all the sub-libraries. The correlation analysis was then repeated with SAGE libraries created elsewhere and of different source material to determine the similarity of the NHMC libraries to independently sampled libraries.

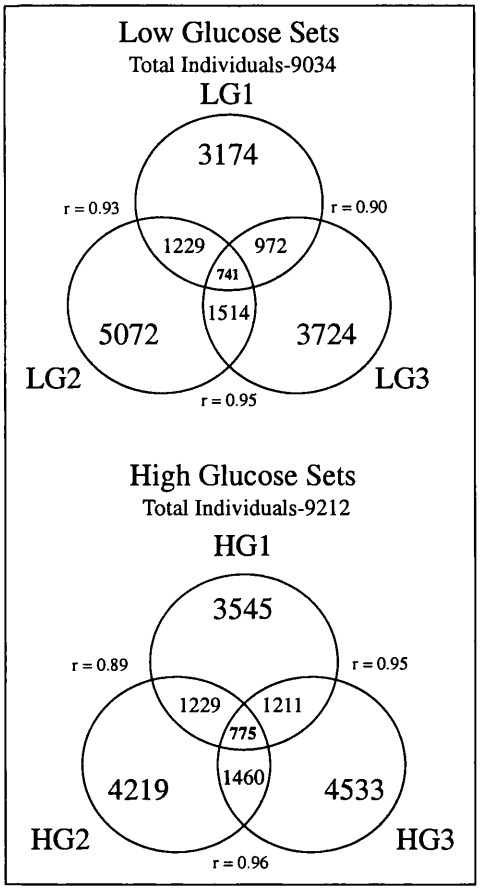


FIGURE 4.2. VENN DIAGRAMMES ILLUSTRATING THE INTERSECTION OF UNIQUE TAGS.

The level of tags that are present in the combinations of sub-libraries. Total numbers of tags are indicated by the largest number in each sub-set. Tags that were present in the intersection of each set of two sub-libraries were used to calculate the Pearson product moment correlation coefficient 'r'. When the intersection of all libraries was used the results changed little.

4.3 COMPARISON OF SAGE LIBRARIES FROM OTHER CELLS AND TISSUES

While data from this analysis would be more different, because cells of different origin and culturing were expected to have different transcriptional profiles, it seemed important to demonstrate that unrelated SAGE libraries were indeed different from each other. The library information is presented in TABLE 4.3a, and calculated correlation coefficients presented in TABLE 4.3b. Note that firstly, all the NHMC SAGE libraries are markedly different from the other SAGE libraries as the other libraries from each other. Secondly, there is a high degree of similarity between all the NHMC libraries as there is for other related libraries such as fibroblasts and astrocytes.

a

Library Abbreviation	Source Tissue/Cell	Description
LGx/HGx	Kidney, Mesangial Cell	NHMC normal (LG) and high (HG) d-glucose
293	Kidney, Embryonic cell line	Embryonic cell line 293, uninduced cells
Cere	Brain	Autopsy sample of cerebellum
Fibro	Fibroblasts	Large T antigen transformed human fibroblasts
H216	pancreas	Normal duct epithelial cells
Heart	Heart	Human normal, bulk tissue
Kid	Kidney	Human normal, bulk
Liver	Liver	Human normal, bulk
NHA	Astrocytes	Normal human astrocytes, cells harvested at passage 5
PR317	Prostate	Normal prostate, microdissected
Pros	Prostate	Normal prostate tissue, epithelium and stroma

TABLE 4.3 A & B. COMPARING NHMC LIBRARIES TO INDEPENDENT LIBRARIES.

SAGE libraries used for the calculation of correlation coefficients (a). All libraries are available through the NCBI web site (ncbi.nlm.nih.gov). Pearson product moment coefficients for each of the 60 correlations (b). Note that each NHMC sub-library is highly similar to each other and markedly different from each of the independent libraries. Each of the independent libraries has low correlation to other SAGE libraries. Independent SAGE libraries were obtained (ncbi.nlm.nig.gov/SAGE). Cell values in bold indicate substantial similarity to NHMCs

b

Pearson r	LG1	LG2	LG3	HG1	HG2	HG3	293	Cere	Fibro	H216	Heart	Kidney	Liver	NHA	PR317	Prostate
LG1	1	0.92	0.89	0.91	0.91	0.91	0.4	0.22	0.47	0.38	0.28	0.2	0.08	0.58	0.54	0.24
LG2		1	0.93	0.91	0.92	0.93	0.45	0.26	0.52	0.45	0.31	0.21	0.09	0.62	0.57	0.27
LG3			1	0.88	0.9	0.9	0.47	0.25	0.53	0.46	0.29	0.2	0.09	0.62	0.57	0.27
HG1				1	0.91	0.91	0.41	0.24	0.49	0.41	0.27	0.18	0.08	0.6	0.53	0.24
HG2					1	0.93	0.43	0.25	0.48	0.4	0.29	0.2	0.09	0.61	0.56	0.25
HG3						1	0.45	0.28	0.51	0.44	0.31	0.22	0.09	0.62	0.57	0.26
293							1	0.29	0.52	0.66	0.25	0.21	0.09	0.47	0.53	0.3
Cere								1	0.28	0.32	0.46	0.61	0.1	0.42	0.37	0.19
Fibro									1	0.57	0.31	0.26	0.11	0.5	0.52	0.47
H216										1	0.22	0.2	0.07	0.41	0.46	0.37
Heart											1	0.46	0.14	0.37	0.55	0.22
Kid												1	0.11	0.43	0.35	0.15
Liver													1	0.09	0.15	0.07
NHA														1	0.53	0.19
PR317															1	0.35

Interestingly two libraries derived from prostate, 'PR317' and 'Pros' appear unrelated. The library data was unclear as to the exact nature of the micro dissection used to create PR317. With prostate tissue being histologically rich in smooth muscle cells, which are closely related to MC and 'Pros' being composed of 30:70 epithelial cells and stroma, it is possible that the PR317 samples were probably rich in smooth muscle. This would account for the similarity between the NHMC and PR317 libraries.

An alternative, graphical representation is presented in FIGURE 4.3, where the standardised frequencies for each matched pair are used to calculate a 'linkage distance' between each SAGE library. The linkage distance is then used to connect each library based on similarity. The similarity between NHMC sub-libraries is reflected in the tight clustering of each sub-library in the same branch and a smaller linkage distance. Additionally, similarities between other libraries are seen as clustering on closer sub-branches but larger linkage distance from NHMC libraries. Dissimilarity between NHMC libraries and other SAGE libraries is seen as still further linkage distance between the tree branches, such as those of the liver and bulk kidney SAGE libraries.

4.4 COMBINING SUB-LIBRARIES

Correlations are useful for determining the degree of similarity between two populations. From the data presented above it seems likely that each sub-library is highly similar to each other and, as derived from the same culture conditions, can be combined to form the primary SAGE libraries. To combine the sub-libraries each sequence file was re-loaded into the SAGE analysis and the entire set of some 1,500 files for each of HG and LG libraries was re-analysed. This was necessary as the phenomenon of duplicate ditags had to be accounted for across the whole of the data for each library, rather than the sum of the single libraries. In practice, the data differed little from the sum of the sub-libraries.

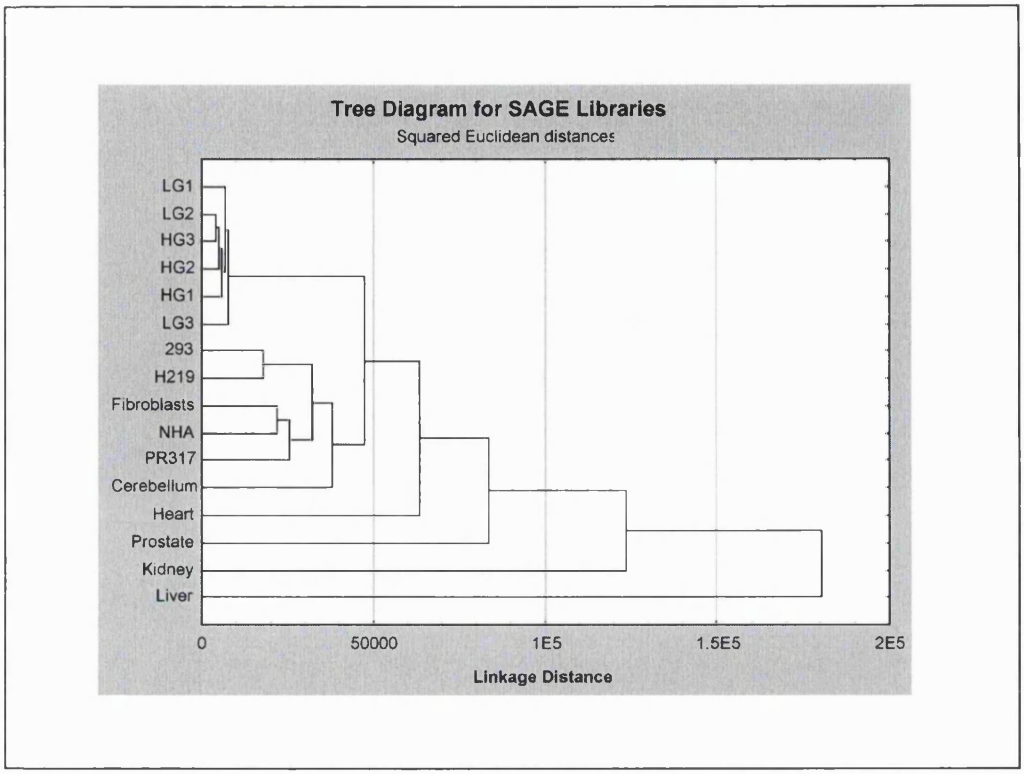


FIGURE 4.3. GRAPHICAL REPRESENTATION OF LINKAGE BETWEEN SAGE LIBRARIES.

Each SAGE library is joined to it's nearest neighbour in a simple clustering of Euclidean distance. The tighter the linkage distance then the more related the library. As can be seen each NHMC sub-library is very similar, while the next most related libraries derive from fibroblasts, smooth muscle (PR317) and astrocytes. Interestingly, the cerebellum library also clusters with the next most related, reasons for this are unclear. The most dissimilar libraries derive from liver, kidney and prostate.

4.5 TAG FREQUENCY DISTRIBUTION AND PROBABILITY OF DETECTION

4.5.1 FREQUENCY DISTRIBUTION

Once combined, general statistics for the sub-libraries were determined. A similar pattern of abundance was observed compared to the sub-libraries and indeed other SAGE libraries. The bulk of the tag mass was represented by few individuals (1.6%), while the majority of unique tags were present only once (see TABLE 4.4).

SAGE Library	Distribution				Totals
	> 20	19 to 5	4 to 2	1	
Unique Tags	241	1033	3278	10401	14953
No. of Tags	16033	9263	8291	10401	43358
% Individuals	1.6	7	22	70	
% In Population	37	21	19	24	100
% Individuals in Pop	1	2	8	24	35

TABLE 4.4. TAG DISTRIBUTION IN COMBINED LIBRARIES.

In order to account for PCR bias, all sequence files were loaded into two SAGE analyses for each of low and high glucose. Linkers were removed and the tag frequencies were grouped in abundance classes. As with other SAGE libraries the majority of tags were present at low abundance while the bulk of tag mass was represented by only a few individuals. Note that the totals vary to a small degree from the sum of the values of the sub-libraries; this is due to the sequence files being analysed again as a single project and duplicate di-tags removed across each project library. Note also that distribution is based on relative abundance and so the classes differ slightly from distributions from previous sets. They are, however, relative estimations based on increases in sampled populations.

4.5.2 DETECTING A TRANSCRIPT

The probabilities of detecting a particular transcript of a particular abundance were calculated (see TABLE 4.5), based upon a mathematical calculation of sampling complex populations described in (Chen et al., '98a) (see CHAPTER 2.8.2). A more thorough description of sampling evaluation is assessed using Monte Carlo simulations and complex hypothetical transcriptomes, which are described and fitted to a more accurate mathematical equation (Stollberg et al., '00).

The data suggest that there is high probability of detecting genes that are present at an abundance of at least 0.03%. Clearly, the NHMC SAGE libraries are not complete, with only some 14% of tags below 0.0025% detected, yet the sampling level presented here should represent a significant proportion of the higher abundances of the NHMC transcriptome.

Abundance (median n)	p(Detection in NHMC)
0.0025% (1)	0.14
0.0075% (3)	0.35
0.03% (12)	0.84
>0.05% (>20)	0.99

TABLE 4.5. PROBABILITY OF DETECTION.

The probability that a tag of a particular abundance class is detected in this library. Calculations are determined using the median value of each abundance class using the combined libraries (TABLE 3.10).

4.6 EXPERIMENTAL ERRORS

As with any analyses involving complicated experimental protocols and the sampling of complex populations, there are several areas where errors can accumulate. Experimental errors will be addressed in this section. Potential (or estimated errors) regarding sampling and mapping, for which empirical data cannot be easily collected, will be addressed in the final section of this chapter (CHAPTER 4.7). Three types of error in the generation of SAGE libraries can be measured, the efficiency of tag generation, the contamination of tag populations by linker sequences or digestion products and the bias caused by PCR amplification. When each of these errors was investigated in this SAGE project, the results indicated a high degree of robustness in the experimental protocol.

4.6.1 EFFICIENCY OF TAG GENERATION

The generation of SAGE tags can be a major source of error. Mapping tags to genes, described more fully in CHAPTER 5, is dependent upon the accurate application of the SAGE criteria, i.e. that the SAGE tag is derived from the position in the transcript defined by the anchoring enzyme, in this project N*la* III (CATG$^{\vee}$). The nature of restriction enzyme digestion in the context of the SAGE protocol means that if digestion is inefficient, then the position of SAGE tag generation will be compromised. If the point of tag generation is not constant for all transcripts, then the identification of individual transcripts will be flawed.

In order that the error in generating SAGE tags could be measured, an experiment was conducted where potential tags from all the N*la* III sites were artificially extracted from a number of Genebank reference sequences and the frequency of these tags in the NHMC sub-library was determined. A graphical representation of tag generation is presented in FIGURE 4.4.

Reference sequences were downloaded from NCBI, and then up-loaded into a sequence editor (EditSeq, DNAstar) where the anchoring enzyme sites were identified. After confirming the orientation of the sequences the 10bp tags, 3'to the AE sites, were extracted and compiled into a database table. This process was repeated for all the candidate genes and the frequency of all the tags in the NHMC sub-library was generated (TABLE 4.5).

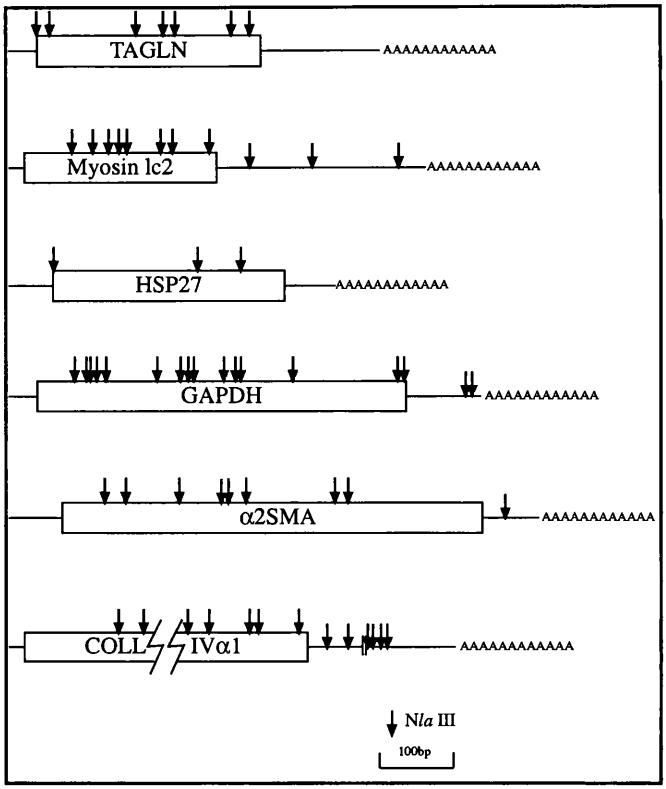


FIGURE 4.4. REPRESENTATION OF THE POSITION OF SAGE TAGS.

Each gene is drawn to scale in conventional 5'-3' orientation. The poly-A tail is noted and the AE (*Nla* III) sites for the entire genes are represented as an arrow (↓). Tags were extracted for each of the sense (3') and anti-sense (5') direction for the analyses. Genes shown are Transgelin (TAGLN), Myosin light chain isoform 2 (Myosin lc2), Heat shock protein 27 (HSP27), Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), Actin-α2, smooth muscle (α2-SMA), and Type IV collagen subunit α1 (COLL IVα1).

From the data, (TABLE 4.5) it can be seen that for the majority of genes the most 3' tag was indeed the major tag in the NHMC sub-library with an average efficiency of 94%, i.e. 6% of tags failed to be generated from the most 3' anchoring enzyme site. Furthermore, when all the artificial tags were mapped using SAGEmap, only 7 mapping references were to genes not related in some way to the primary tag suggesting that the efficiency of tag identification was even higher. There also appeared even expression levels across the sub-libraries, which supports the linear generation of tags across each sub-library thus reproducible experiments.

Gene (UniGene cluster)	Acc No. Length bp	Artificial Tag Position (bp)	Frequency in Sub Libraries (Total Tags)						Freq in Full NHMC Library	Mapping (UniGene ID)
			HG1 6346	HG2 7599	HG3 8549	LG1 5662	LG2 9924	LG3 6795		
Transgelin (Hs.75777)	NM_003186 1085bp	655 (1°)	86	97	97	89	123	70	505	Hs.75777
		605 (2°)	0	2	1	0	0	1	3	Hs.75777 Hs.16089 Hs.49587
		454 (3°)	1	2	0	0	1	0	4	Hs.75777
Myosin lc (Hs.9615)	NM_006097 1120bp	1048 (1°)	36	41	45	29	55	26	214	Hs.9615
		820 (2°)	1	2	1	0	2	1	7	Hs.9615 Hs.236545
		653 (3°)	0	0	0	0	2	1	3	Hs.9615
		548 (4°)	1	0	0	1	1	0	2	Hs.9615
		442 (5°)	1	0	0	0	0	0	1	Hs.9615
HSP27 (Hs.76067)	NM_001540 865bp	624 (1°)	8	15	16	14	29	18	94	Hs.76067 Hs.37417
		511 (2°)	2	2	0	1	2	0	7	Hs.37417
COLLIVα1 (Hs.119129)	NM_001845 6447bp	6293 (1°)	43	32	44	28	54	47	225	Hs.119129
		6258 (2°)	2	1	0	0	0	1	4	Hs.119129 Hs.159608
		6038 (3°)	1	0	2	1	1	0	5	Hs.119129 Hs.353639
		5179 (5°)	2	0	3	2	2	2	9	Hs.119129 Hs.115907
		4807 (6°)	0	0	0	1	1	0	1	Hs.119129 Hs.18166
		2977 (14°)	0	1	1	0	0	0	2	Hs.119129
		635 (23°)	0	0	2	0	0	0	2	Hs.119129
αSMA (Hs. 195851)	NM_001613 1330bp	1038 (1°)	82	100	88	69	101	54	445	Hs.195851 Hs.1288
		914 (2°)	0	2	3	2	1	3	11	Hs.195851
		889 (3°)	1	0	2	1	1	0	1	Hs.195851
		591 (5°)	0	0	0	0	1	0	1	Hs.118127
		309 (8°)	0	0	0	1	1	0	1	Hs.269926 Hs.195851
GAPDH (Hs.169476)	NM_002046 1283bp	1263 (1°)	16	28	33	19	35	26	146	Hs.169476
		1215 (2°)	6	2	1	1	1	1	9	Hs.169476 Hs.356772
		1078 (3°)	0	1	2	0	1	0	3	Hs.169476 Hs.356772 Hs.234546

TABLE 4.5. GENE SEQUENCES USED TO GENERATE THE ARTIFICIAL TAGS.

Reference sequences are annotated as NM_ accession numbers. The position of the tag and the hierarchical tag level (in brackets) are shown with the frequency of appearance in the NHMC library. Note that the tag level is not always sequential as the 3rd tag precedes the 5th tag in COLL IVα1 tag hierarchy. Mapping data indicates the UniGene cluster for each tag. Boldface tags have mapped incorrectly and so constitute an error in tag to gene mapping. Different UniGene clusters not in bold face represent different isoforms or genes that are highly similar to the primary gene.

4.6.2 CONTAMINATION BY 5' ANCHORING ENZYME DIGESTION PRODUCTS

In addition to the correct sense tags (3' to the AE site), anti-sense tags (5' to the AE site) were extracted from the reference sequences and used to scan the project (see TABLE 4.6). This provided an estimation of the level of contamination by digestion products in the protocol. This, in turn, indicates any technical failure of SAGE. The average contamination by digestion products was determined to be 0.8% which again suggests robust experimental technique.

Gene	Tag Sequence	AE Site	Position sense/ antisense	Sub Library Freq.						Freq	Mapping
				LG 1	LG 2	LG 3	HG 1	HG 2	HG 3		
TAGLN	CTTTCCCTCC	2°	605/504	0	0	0	1	0	0	1	Hs.75777
MYOlc	GTGGTGAGCA	6°	442/702	0	0	0	0	0	1	1	Hs.180884
	TTGAAAGCCT	12°	172/972	0	0	0	1	0	0	1	Hs.323567 Hs.9615
HSP27	GGGGCCTCCA	1°	624/265	0	1	0	0	0	0	1	Hs.181165 Hs.371812
COLLIV α1	AGAAACCACG	2°	6258/213	0	1	0	0	1	0	2	Hs.40719
	AACCTGGGTT	4°	6010/461	0	1	0	0	0	0	1	Hs.191979 Hs.33756
	GGCTCAGGGG	20°	2653/3818	0	0	0	0	1	1	2	Hs.4953
αSMA	GTGCTGGGTG	1°	1038	1	0	0	1	1	0	3	Hs.195851 Hs.4209
	GATGCCAGCA	3°	316	0	0	0	1	0	1	2	Hs.195851

TABLE 4.6. CONTAMINATION BY ANTI-SENSE TAGS.

Tags were generated from the reverse complement of the reference sequences described above. The relative positions of the tags present in the NHMC library are shown together with their presence in all the sub libraries. Ambiguous mapping is noted as normal type UniGene cluster ID's while correct mapping is indicated and boldface text. The AE site represented the hierarchical site for correct tag generation

4.6.3 CONTAMINATION BY LINKER SEQUENCES

Analysis of previous SAGE projects has shown that contamination by linker-derived sequences can be a major source of error. A total of 1254 tags, (2.8%) were removed from our project library constituting the major source of contamination (see TABLE 4.7). This level of linker contamination is consistent with previously published SAGE projects. (Velculescu et al., '95)

Linkers Removed	Total Frequency (abundance)	Frequency in Sub-libraries					
		LG1	LG2	LG3	HG1	HG2	HG3
TCCCGTACA	472 (1.1%)	60	99	56	92	88	77
TCCCTATTAA	411 (1.0%)	63	93	12	53	80	110
TCCCCGACAT	48 (0.1%)	6	20	4	8	3	7
TCCCCGTAAT	67 (0.2%)	12	17	11	10	9	8
TCCCTATTG	18 (0.04%)	1	7	3	1	3	4

TABLE 4.7. EXAMPLES OF LINKER SEQUENCES REMOVED FROM THE SAGE LIBRARY.

The most prominent tags were exact matches to the linker sequences and subordinates that match greater than 8/10 bases were also removed. Even though the subordinate tags were present at a much lower frequency they were checked for potential mapping data prior to removal from the library, and most could not be mapped correctly to a gene.

4.6.4 PCR BIAS

The phenomenon of biased PCR amplification was addressed in the original SAGE projects and generally considered and accounted for in the tag extraction software. Briefly, the likelihood of two of the same tags joining together during the formation of ditags (CHAPTER 2.6.4) has been estimated to be rare (Velculescu et al., '95, Zhang et al., '97). With this in mind, any bias in PCR amplification (seen as increases in one particular ditag) can be accounted for by including only unique ditags in the SAGE library report. This step is included in the SAGE software used to extract the tags from the sequence file. The analysis of duplicate di-tags is summarised in TABLE 4.1. The frequency of duplicates in each library was determined to be 22 ±1.9% for the HG sets and 13 ±1.6 % for the LG sets indicating, that while there was a difference in contamination between the two experimental groups, within each sub-library the variation from the mean was not great.

4.7 DISCUSSION

Following the validation of the culture protocol and feasibility of the SAGE technique a series of SAGE libraries were generated from independent cultures and sampled to the level of 45,000 tags. Approximately one half were from cells grown under normal glucose and one half from cells grown under high glucose. The sampled libraries were analysed for frequency distribution and compared to SAGE libraries generated from other studies. The SAGE libraries generated here showed similar frequency distribution of tags to other libraries and high abundance tags were more than likely generated from high abundance genes. The SAGE libraries were also analysed

for experimental errors, which were found to be low, in that 96% of a sample of tags mapped to the correct genes.

Six SAGE sub-libraries were sampled in this project, each from an independent culture series using normal human mesangial cells at the same level of passage (three libraries each from a low and high glucose culture). Initial analysis showed that each of the three libraries was similar for contamination by linker sequences and in the frequency distribution of tags. Each library also displayed a relatively stable generation of tags, indicated by approximately the same level of a particular tag in each sub-library. Both these observations support the robustness of the SAGE protocol and the similarity between the libraries.

Each sub-library was compared to other NHMC sub-libraries and to SAGE libraries produced elsewhere of different source material. These experiments demonstrated two main results that in turn supported the hypothesis that a SAGE library describes a cellular phenotype. Firstly, each NHMC sub-library showed a very high degree of similarity to each of the other sub-libraries (as assessed by a Pearson product moment correlation) and less similarity to other SAGE libraries. Secondly, the NHMC libraries showed a higher degree of similarity between SAGE libraries generated from related cells, such as astrocytes, fibroblasts and smooth muscle cells, than to libraries of un-related cells such as cardiac muscle cells, hepatocytes and bulk kidney tissue. This second observation offers a method of classifying the relationship between cells and their linage at a resolution not previously applied, defined by a global transcriptional profile rather than antigenic markers.

Based on the similarity between the six sub-libraries, each set of three was combined to form the primary experimental libraries for this project. As expected, general library statistics were in agreement with other SAGE libraries of similar magnitude and both the NHMC-LG and NHMC-HG libraries were remarkably similar.

With respect to an accurate measure for the experimental generation of the NHMC libraries, empirical data was determined for the efficiency of tag generation, the contamination by linker and anti-sense tags and the bias caused by PCR. All three were determined to be of low magnitude and acceptable. Estimating other errors in such a project is more theoretical and will be discussed below. A first step in addressing the

non-empirical errors in a SAGE library is to assess the errors involved in tag sampling. Other errors that cannot be directly measured include sequencing errors, non-randomness of DNA and the non-uniqueness of tags. All these errors, together with those presented in section 4.6, will combine and compromise the accuracy of a SAGE library.

The frequency distributions of sampled SAGE libraries appear to form a particular function. These distributions are characterised by the vast majority of transcripts being of low frequency and the bulk of transcripts being composed by a few individuals. Previous mRNA abundance studies indicate discrete abundance classes (Bishop et al., '74, Hastie and Bishop, '76), but it is likely that these abundance classes are specific for a particular transcriptome. Many SAGE libraries are described in abundance classes and the assignation of abundance classes appears almost arbitrary. One of the largest SAGE analyses involved colonic epithelial cells (Zhang et al., '97) where the frequency ranges for transcript copy number per cell were 1-5 (64%), 6-50 (31%), 51-500 (4.38%) and >500 (0.42%). To assume a stepwise function with discrete ranges is simplistic and essentially inaccurate. This assignation would imply that a tag present at a frequency of 500 has a far lower probability of detection than a tag present at a frequency of 501, as its abundance is implied to be some ten-fold lower. This is clearly not the case and so alternatives are used. To date, the only method for estimating tag sampling is to use a double exponential function matched to the ranges empirically determined from other SAGE projects (Stollberg et al., '00). This model has proved useful in estimating the sampling error in SAGE projects and consequently the level of sampling required for accurate library generation.

The sampling of complex populations is problematic. The degree to which a sample can represent the population and the difference to the real population is the error associated with sampling. Investigators have used two hypothetical transcriptomes to assess the sampling error, 15,000 and 78,000 (Stollberg et al., '00). The transcriptome of 15,000 was sampled effectively with 11,460 unique tags (to 98.5% as an estimation of detection) while the transcriptome of 78,000 dropped to 92.5% when sampling 64,000 unique tags. The real human genome should lie between these two estimates as the estimate of genes is currently between 30,000 and 40,000, yet many SAGE projects report the continued detection of unique tags beyond of 60,000(Madden et al., '97). Conservative estimates have suggested that 650,000 tags are required to adequately

sample a transcriptome of 56,000 (Velculescu et al., '00). This represents a very large project and a definition of 'adequate sampling' is absent.

Sequence errors can result from two phenomena; errors in PCR amplification and errors in single pass sequencing. Mis-incorporation of nucleotides in PCR amplification of 100bp fragments is considered insignificant with the current T*aq* polymerases. Single pass sequencing errors have been calculated to be between 0.7-1% per base in this protocol (Applied BioSystems Prism). In a SAGE library of 40,000 ten base pair tags this corresponds to about 4,000 tags. The error, when it occurs, will be relevant as it will lower the correct tag count for a transcript by one and increase an alternate transcript by one. The error may also create an entirely new tag. Because sequencing errors are random events, each tag has an equal probability for such an error. Tags with high counts will be the most robust as the removal or addition of a single tag will not alter the relative frequency significantly, thus these errors will be of little importance. More care is required when examining tags that are present at lower levels as artificially inflated or deflated frequencies may occur. This random nature of sequencing errors suggests that the vast majority of these errors will be in tags present at a frequency of 1, and will erroneously map to a gene or not map at all. Several methods to address these problems have been suggested, but currently it is convention to remove all single tags. While this may not be optimal there appears no clear method for eliminating sequencing based errors and in any case tags present only once offer a low data value.

DNA is not a random sequence of bases. Because some DNA sequences are more common than others, this will introduce the potential for ambiguous mapping and thus reduce the number of transcripts with unique tags. Experiments using Monte Carlo simulations have been useful in determining the significant of SAGE tag generation and the errors inherent in sampling complex systems. They appear in general agreement that errors exist and are quantifiable using algebraic formulae. Much of these simulations require assumptions regarding the uniqueness of individuals in the population. A longer tag sequence would confer more likelihood of uniqueness, but simulations have determined that even the use of 10bp tags (potential for 4^{10} individuals) will result in minor reduction in non-uniqueness(Stollberg et al., '00).

Other features of dynamic transcriptomes are difficult to introduce into mathematical equations. The distribution of transcribed pseudogenes, polymorphisms and shared exons will all add to the non-randomness of DNA and thus present potential sources of error. The characterisation of mRNA and its relationship with a gene should provide an indication for this error, but will probably only serve as a measure rather than a mechanism for discrimination and will require scrutinising each relationship. Resolving the transcriptional properties of all genes with regard to the molecules they produce seems incomprehensible with the technology available currently. The nature of transcriptome analysis relies on abandoning, temporarily, the scrutiny of a single gene in favour of obtaining global transcription characteristics. Additionally, the technology of SAGE offers insights into the heterogeneity of transcription at a resolution not previously available to the average researcher and so is useful in other ways.

As well as redundancy in tags, it is not true that a single tag represents a single gene. Even with the estimation of 4^9 or 4^{10}, which far exceeds the number of genes present in the genome, the error of tag uniqueness is apparent. These errors have been estimated previously and are currently believed to result from determining the probability that a tag sequence is unique to one gene, rather than whether a gene is represented by one tag. Here they are estimated at 1.5% for 10bp tags and a 15,000 genome and 7.2% for 10bp tags and a 78,000 genome. An empirical measure for this SAGE library will be assessed in CHAPTER 5.

In summary, six SAGE libraries were successfully constructed from an *in vitro* model of DN and sampled to a high level. Each sub-library was similar with regard to frequency distribution of tags and the complexity of those tags. The technical errors in generating the libraries were low, with the generation of tags from non-primary AE sites (6%) and random sequencing errors (7-10%) the largest. The sampled libraries were compared to other SAGE libraries and found to more similar to each other than to independent libraries of other sources. Additionally, NHMC libraries were more similar to libraries from cells of similar lineage than to un-related cells and thus can be thought of as a classification system based on global transcription.

CHAPTER 5

5 GENERATION & VALIDATION OF THE NHMC TRANSCRIPTOME

5.1 INTRODUCTION

Sampling the SAGE libraries in CHAPTER 4 demonstrated that the frequency distribution of tags was similar for other SAGE libraries, which indicated a complex population. Yet the sub-libraries were more similar to each other than unrelated SAGE libraries, which suggests that while the frequency distributions are similar the individual tags are not. This led to the idea that a transcriptome could be used to classify cell types, which was demonstrated by comparing independent SAGE libraries from a variety of sources. These comparisons showed that the cells of similar lineage had more similar transcriptomes than unrelated cells. The experiments in this chapter describe the generation and validation of the NHMC transcriptome by identifying which genes are present, at what level and whether they can be rationalised in the context of the NHMC. While 'mapping' genes has classically referred to the physical placement of genes in the genome, SAGE mapping refers to the assignment of tags to genes, or more accurately to UniGene clusters. The continued anchoring of UniGene clusters to genomic locations (loci) means that SAGE data may also be used to physically map genes.

The data product of a SAGE analysis is a list of tags together with a frequency of appearance. The basis of analysis is that this digital output represents the transcriptional profile of the cell. There are two problems associated with a SAGE analysis, which result in a loss of fidelity between the actual transcriptional profile and the SAGE derived profile. Accuracy in both the assignment of tags to genes and the ability to quantify a gene's expression are sacrificed in order to increase throughput and therefore speed of analysis.

Considering the first problem of tag assignment requires a measurement of the ability of a tag to distinguish between genes. The estimation of errors described in the previous chapter demonstrates that this process is difficult and results in a reduced accuracy in mapping. Currently, mapping of tags to genes takes the form of identifying the tag in the nucleic acid databases at NCBI, where approximately 1.5 million sequences are stored. Only 18,000 of these are characterised mRNAs, the majority being ESTs. Assembling all the sequence data into 'gene clusters' is attempted (e.g. UniGene) but this is an evolving dataset and is based primarily on sequence homology so the subtle changes in transcriptional variance, i.e. alternative splicing, premature termination and other transcriptional variations are not immediately apparent.

Considering the second problem of 'data validity', the greatest non-empirical errors have been ascribed to sequencing (as described in CHAPTER 4). However, the assignment of tags will also feature in this problem as incorrect assignment not only increases one genes expression, it will lower another's. Quantitation of gene expression is dependent on the correct assignment of tags to genes.

The experiments described in this chapter address these two problems and describe the assignment (mapping) of tags to genes, then estimate the accuracy of these assignments as a quantitation of gene transcription. A general description of the NHMC transcriptome is presented.

5.2 STRATEGY FOR MAPPING SAGE TAGS TO GENES

Initially, individual tags were subjected to a multi-step procedure of identification that involved placing tag sequence in low complexity BLASTn alignment algorithms through NCBI. The tags were queried across two databases, the 'nr' database that favours the detection of characterised entries and the 'EST' database that favours the detection of all expressed genes. The small size of the alignment sequence (14bp) meant such alignment programs returned hundreds of matches in both the 'nr' and 'EST' datasets. Each set of sequences was then inspected for the official SAGE criteria, namely that the tag was generated from the most 3' N*la* III site of the cDNA sequence. Although this method of analysis was used in the pilot THP-1 project, where relatively small SAGE libraries were generated, it was a time consuming process and impractical for large SAGE libraries. During the course of this project extracting SAGE tags from sequence files and mapping SAGE tags to genes became more straightforward and streamlined. This was due to the advances in software for processing raw sequence data and the availability of tag mapping data.

Recently, extensive mapping information for SAGE has become available from NCBI servers (see TABLE 1.1). These data are simple delimited text files that contain three fields of information, the tag sequence, a UniGene cluster ID and a description of the gene cluster. The datasets are compiled from all submitted SAGE data (currently

some 5,000,000 tags) and tag sequences extracted from UniGene clusters for any combination of anchoring enzyme (AE) and tagging enzyme (TE). Three dataset sets are available for each anchoring enzyme; a reliable 'tag-to-gene' (TTG) mapping set where data consists of characterised mRNA or oriented EST sequences, a complete TTG mapping set where the tags are mapped to all the sequences clustered in UniGene and finally a 'gene-to-tag' (GTT) mapping set where SAGE tags are generated *in silico* from UniGene clusters. Currently, the reliable TTG mapping set contains some 246,000 entries and the GTT set contains 99,000 entries, the difference being an indicator of the accuracy and complexity of characterised cDNA in UniGene clustering together with its insensitivity to alternative transcription products.

Several issues concerning the mapping of SAGE data became apparent during the pilot project and, as the technology has become more widely utilised, from the warehousing of SAGE data. It is clear that there are more SAGE tags than even the most over estimated level of genes in the genome, currently 3-fold redundancy (99,000 compared to 30-40,000 estimated genes). From the initial inspection of pilot SAGE project (CHAPTER 3.1), it was apparent that some tags are not unique to genes and unique tags do not always represent individual genes. These observations belie two phenomena that can compromise the quality of the SAGE mapping data, first the masking of genes by tags and second, the masking of tags by genes. Clearly, this undermines the data in a SAGE library, as although a 10bp sequence should discriminate between 4^{10}, individuals it is not true that the tags generated from mRNA are unique to each gene in this same magnitude. This issue of redundancy will be discussed in CHAPTER 5.

With the available NCBI mapping data, local databases were constructed and used to map the NHMC-SAGE library to UniGene clusters (Lal et al., '99, Lash et al., '00). The datasets were uploaded into Microsoft Access tables and a number of queries used to analyse the data. An initial query mapped the NHMC library to matches in the SAGE database, a second query identified tags that failed to match. Finally a series of filters were used to refine the mapping data into a series of transcriptome maps representative of the levels of redundancy. This process is represented as a flow chart in FIGURE 5.1.

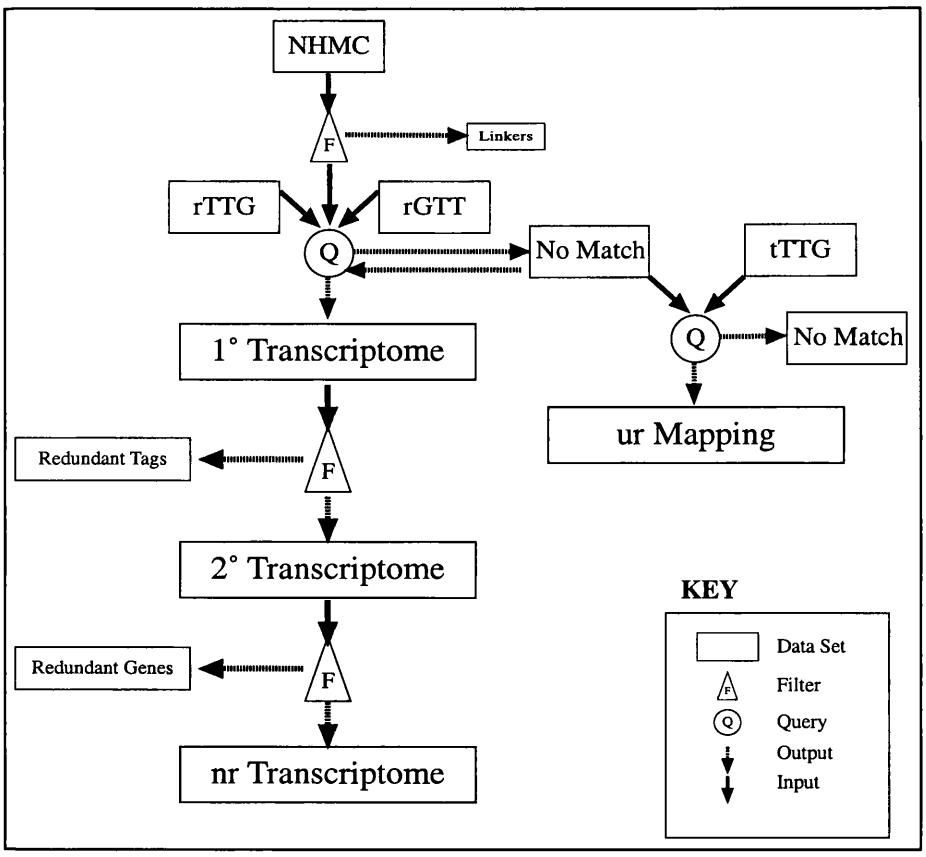


FIGURE 5.1. SCHEMA FLOW CHART OF THE PROCESS USED TO MAP TAGS TO GENES.

The NHMC was initially filtered for linker sequences. The library was then loaded into an MS Access table and the tag sequences present in NHMC that matched entries in the combined tables of 'reliable Tag-To-Gene' (rTTG) and 'reliable Gene-To-Tag (rGTT) was queried. The output formed the 1° transcriptome and tags that failed to match. Un-matched tags were queried in the 'total Tag-To-Gene' (tTTG) table and an unreliable mapping set was generated. The 1° Transcriptome was then filtered for redundant tags creating the 2° Transcriptome. The 2° Transcriptome was further filtered for redundant genes producing the non-redundant (nr) Transcriptome. No match tags were periodically re-queried in the reliable database and as the resource at SAGEmap grew.

The primary transcriptome contains raw mapping data and ignores the redundancies inherent in the SAGE analysis. This transcriptome therefore overestimates the number of genes present in the real transcriptome. The first filter applied to the primary transcriptome removes the redundant tags i.e. tags that map to more than one gene. This creates the secondary transcriptome that, although somewhat smaller, is a more accurate measure of the genes present in the cell system. The next filter, applied to this secondary transcriptome, removes the redundant genes, i.e. all the genes that map to more than one tag. This final filter creates the non-redundant transcriptome where one tag represents one gene and one gene is represented by one tag, the most stringent criteria.

5.2.1 GENERATION OF THE PRIMARY NHMC SAGE TRANSCRIPTOME

For the purposes of this chapter the data for each library (low glucose, NHMC-LG), and high glucose, NHMC-HG), is combined into one NHMC library. The local mapping databases contained information for some 246,000 tags derived from the NCBI TTG and GTT mapping databases. When combined, the NHMC SAGE library consisted of 43,358 tags, representing 14,953 individual tag sequences. When applied against the local SAGE database the output returned matches for 20,382 tags and failed to match 4,116 tags. This represents the raw mapping data from this SAGE project or the primary transcriptome.

From the disparity between the total number of unique tags and matches returned, it was apparent that both types of masking effect were present. Using a sample of the top 1000 tags from the primary transcriptome, 79% matched a single UniGene entry, 16.9% matched between 2 and 4 entries and 4.1% matched more than 4 UniGene entries. Clearly, the mapping data from the primary transcriptome is problematic. It is possible to resolve some of this tag redundancy by including the 11^{th} bp in the mapping procedure. Previous SAGE libraries have used this to resolve redundant tags, and this method is used to filter candidates for differential transcription experiments in CHAPTER 7.

Examining the mapping data from the top 100 tags revealed a predictably high frequency of tags for genes associated with the cytoskeleton, such as actin, transgelin, myosins and protein synthesis genes such as the ribosomal proteins. Other genes present at high abundance were associated with metabolism such as GAPDH and a variety of transcription and translation factors. TABLE 5.1 lists the top 50 tags from the primary transcriptome together with mapping data.

Tag Sequence	Total count	UniGene ID (Hs.)	Gene Description
ACAGGCTACG	505	75777	transgelin
AAGATCAAGA	445	-	3 matches (actin)
ATGTGAAGAG	389	111779	SPARC (osteonectin)
GTGCTGAATG	302	77385	myosin, lp 6, alkali, smooth muscle
TTTGCACCTT	245	75511	CTGF
ATCTTGTTAC	239	287820	fibronectin 1
CACAAACGGT	238	195453	RPS27 (MPS 1)
TTGGTCCTCT	234	-	2 matches
GACCGCAGGA	225	119129	collagen, type IV, alpha 1
GGAGTGTGCT	214	9615	myosin lc 2, smooth muscle isoform
CATATCATTA	212	119206	IGFbp 7
AAGACAGTGG	185	5566	RP L37a
TTGGTGAAGG	182	-	2 matches
TTCTGTGAAT	178	-	3 matches
AAGGAGATGG	175	-	2 matches
CCCATCGTCC	166	mito	mitochondrial sequence
GCCCCCAATA	165	227751	lectin 1 (galectin 1)
TGTGTTGAGA	161	-	2 matches
AAAGTCATTG	147	-	2 matches
TACCATCAAT	146	169476	GAPDH
CCTGTAATCC	139	-	131 matches
AGGCTACGGA	135	119122	RP L13a
TGCCTCTGCG	135	75564	CD151 antigen
AGCACCTCCA	131	75309	translation EF 2
ATTCTCCAGT	128	-	2 matches
CCAGAACAGA	125	-	3 matches
TAGGTTGTCT	125	-	2 matches
CCGTGACTCT	124	296267	follistatin-like 1
GACCAGGCCC	124	300772	tropomyosin 2 (beta)
GGGCTGGGGT	122	-	2 matches
GCATAATAGG	117	184108	RP L21
GTGAAACCCC	115	102178	127 matches
CGCCGCCGGC	114	182825	RP L35
CCACTGCACT	112	-	74 matches
GTTGTGGTTA	107	75415	beta-2-microglobulin
TCAGATCTTT	105	108124	RP S4, X-linked
CTGGGTTAAT	105	298262	RP S19
ATAATTCTTT	101	-	2 matches
ACATCATCGA	98	182979	RP L12
CCTAGCTGGA	98	-	2 matches
TTTGGTTTTC	95	179573	collagen, type I, α2
CCCAAGCTAG	94	-	2 matches
CCCGTCCGGA	93	180842	RP L13
AAGGTGGAGG	91	-	2 matches
TAAGGAGCTG	90	299465	RP S26
GACGACACGA	90	153177	RP S28
TTAAAGATTT	84	77899	tropomyosin 1α
TGGTGTTGAG	82	275865	RP S18
GTTTATGGAT	82	279009	matrix Gla protein
GCCGAGGAAG	80	-	2 matches

SUMMARY Top 1000	
Total Tags	1000
Reliable Matches	832
Unreliable Matches	168

TABLE 5.1. PRIMARY NHMC TRANSCRIPTOME (TOP50).

As predicted there is a high abundance of genes associated with the cytoskeleton, translation and metabolism. Approximately 17% of the top 100 tags returned multiple matches, and thus were deemed unreliable without further characterisation

5.2.2 CONDENSING THE PRIMARY NHMC TRANSCRIPTOME INTO THE SECONDARY TRANSCRIPTOME

From the primary transcriptome it appeared that there was a large amount of redundant tag mapping, i.e. unique tags that map to more than one gene. A total of 8,529 tags were removed from the primary transcriptome leaving 11,853 tags constituting the secondary transcriptome. A large volume of the redundant tags appeared to be rich in adenine residues, which suggest an anchoring enzyme site very close to the poly-A tail or other regions rich in adenine residues. Regions of the mRNA rich in adenine may well serve as incorrect priming sites for cDNA synthesis and thus constitute a further source of experimental error. Some tags appeared not to be rich in any particular base and as such may be genuinely generated and represent non-unique sequence or other conserved regions in a transcript. It would be interesting to investigate whether these redundancies do indeed represent conserved regions, shared introns, signal sequences or any other marker that would group these otherwise seemingly un-related genes.

Of the 11,853 matches in the secondary transcriptome, 79% matched characterised cDNA entries, 16% matched EST entries, and 5% of tags remained unmatched (see TABLE 5.2). A second level of redundancy existed in the SAGE library where some 26% of tags mapped to the same UniGene entry. The top 50 tags and genes did not alter significantly from the primary transcriptome (TABLE 5.1), in that genes associated with the cytoskeleton and ribosomal proteins are still present. This implies that high abundance genes are more likely to be highly characterised and thus have more accurate mapping data. The top 200 tags of the 2° transcriptome are presented in APPENDIX 4.

Tag Sequence	Total	UniGene ID	Gene Description
ACAGGCTACG	562	75777	transgelin
ATGTGAAGAG	424	111779	SPARC (osteonectin)
CACAAACGGT	266	195453	RP S27 (MPS 1)
TTTGCACCTT	260	75511	CTGF
ATCTTGTTAC	255	287820	fibronectin 1
GACCGCAGGA	248	119129	collagen, type IV, α 1
CATATCATTA	232	119206	IGFbp 7
GGAGTGTGCT	232	9615	myosin lc 2, smooth muscle isoform
AAGACAGTGG	199	296290	RP L37a
TTGGTGAAGG	192	75968	thymosin, β 4, X chromosome
CCCATCGTCC	182	mito	Tag matches mitochondrial sequence
GCCCCCAATA	181	227751	Lectin 1, (galectin 1)
TACCATCAAT	157	169476	GAPDH
TGCCTCTGCG	151	75564	CD151 antigen
AGGCTACGGA	149	119122	RP L13a
AGCACCTCCA	143	75309	Translation EF 2
GACCAGGCCC	138	300772	Tropomyosin 2 (β)
ATTCTCCAGT	136	234518	RP L23
CCGTGACTCT	131	296267	follistatin-like 1
GCATAATAGG	125	184108	RP L21
CGCCGCCGGC	123	182825	RP L35
GTTGTGGTTA	115	75415	beta-2-microglobulin
CTGGGTTAAT	114	298262	RP S19
TCAGATCTTT	112	108124	RP S4, X-linked
ACATCATCGA	105	182979	RP L12
CCCAAGCTAG	100	76067	HSP 27kD 1
AAGGTGGAGG	98	163593	RP L18a
CCCGTCCGGA	97	180842	RP L13
TAAGGAGCTG	97	299465	RP S26
GACGACACGA	97	153177	RP S28
TGGTGTTGAG	91	275865	RP S18
GTTTATGGAT	88	279009	matrix Gla protein
GGATTTGGCC	87	351937	RP, large P2
GAGGGAGTTT	83	76064	RP L27a
TTCATACACC	83	mito	Mitochondrial sequence
TTGGGGTTTC	82	62954	Ferritin, heavy polypeptide 1
TGCATCTGGT	79	75410	HSP 70kD 5
ACAGATTTGA	77	41271	EST, EUROIMAGE 1913076
GTTCGTGCCA	77	287361	RP L35a
TTACCATATC	77	300141	RP L39
TTGTTGTTGA	73	182278	calmodulin 2
GGGGAAATCG	73	76293	thymosin, β 10
CAATAAATGT	72	337445	RP L37
GGCTGTACCC	71	108080	cysteine and glycine-rich protein 1
GATGAGGAGA	67	179573	collagen, type I, α2
AGCCCTACAA	67	mito	Mitochondrial sequence
TCCCCGTAAT	67		No match
ACTTTTTCAA	67		No match
AGGGCTTCCA	66	29797	RP L10
TAATAAAGGT	66	151604	RP S8

SUMMARY Top 1000	
Total Tags	1000
Reliable Matches	950
Unreliable Matches	50

TABLE5.2. SECONDARY NHMC TRANSCRIPTOME (TOP50).

While not as problematic as tag redundancy, the phenomenon of gene redundancy may represent alternative transcription of a gene or alternative processing of a gene or the difficulty in cDNA cloning in particular mRNA stability. Note that when the unreliable tag matches are removed from the 1° transcriptome the top 1000 tags are sampled from the revised or 2° transcriptome

5.2.3 GENERATION OF THE FINAL, NON-REDUNDANT TRANSCRIPTOME

The secondary transcriptome was filtered for redundant UniGene cluster ID's. Removal of 3,080 entries revealed the final non-redundant transcriptome. This somewhat diminished transcriptome of 8,774 entries represents the most stringent mapping of the SAGE data where a single tag maps to a single gene (see TABLE 5.3). The set of redundant UniGene entries is not as problematic a data set as the redundant tags. While it would be time consuming to discriminate between genes that map to the same tag, the sets of tags that map to the same gene can contain information on the transcription of a gene. A UniGene cluster that is represented by more than one tag may well indicate alternative transcription. The literature is rich with examples of genes that are alternatively transcribed or processed, and analysis of tags generated in a SAGE project will offer insights into the transcription of genes. An added use for this data is the estimation of efficiency of the SAGE technique. Clearly, tags not generated from the most 3' N*la* III site could also indicate inefficient digestion of the cDNA by the anchoring enzyme. This would constitute a serious technical error and invalidate the SAGE project. This issue was discussed and tested in CHAPTER 4.6.1, and it was determined that the efficiency of the SAGE protocol was high, corresponding to 6% of tags potentially generated from a non-primary AE site.

5.2.4 SUMMARY OF MAPPING DATA

Mapping the NHMC SAGE library proved more informative than simply assigning tags to genes. The mapping data resolved into 3 transcriptomes (1°, 2° and nr) each representing different levels of resolution. Different information can be derived from each level of the transcriptome, and it would be unwise to focus on only one. The non-redundant transcriptome provides the most accurate mapping, but possibly only represent a fraction of transcribed genes. The secondary transcriptome potentially contains information regarding different transcripts that are generated from the same gene. Finally, the primary transcriptome, while unclear with regard actual mapping data, may be useful in identifying conserved regions of genes, whether they are signal sequences, shared exons or other indicators that may be used for clustering otherwise unrelated genes. A summary of the three transcriptomes is presented, (TABLE 5.4.),

Tag Sequence	Total	UniGene ID (Hs.)	Gene Description
GGAGTGTGCT	232	9615	myosin lc 2, SM isoform
TAAGGAGCTG	97	299465	RP S26
GACGACACGA	97	153177	RP S28
ACAGATTTGA	77	41271	EST
GTTCGTGCCA	77	287361	RP L35a
GGCTGTACCC	71	108080	cysteine & glycine-rich protein
TCCCCGTAAT	67		No Match
ACTTTTTCAA	67		No Match
ACTGAGGAAA	58	77326	IGFbp 3
AATCCTGTGG	54	178551	RP L8
ACCTGTATCC	51	182241	interferon induced TransPro 3
CTAAGACTTC	50		No Match
TCCCCGACAT	48		No Match
ATCGTGGAGG	43	727	inhibin, beta A (activin A)
GAGCCTGGAT	42	9004	CSPG 4
CTCAGACAGT	39	108957	40S RP S27 isoform
AGCCTTTGTT	39	9930	SERPIN H 2
CTGGGCGTGT	38	351987	ESTs
TCACCGGTCA	38	290070	gelsolin
GGTTGGCAGG	38	3745	MFG-EGF factor 8
TCCCGTACAT	38		No Match
AGCAGATCAG	33	119301	annexin II ligand
ATCAAGGGTG	32	157850	ribosomal protein L9
GGGCGCTGTG	32	8372	ubiquinol-cyc c redase SU
CTGAGAGCTG	31	78501	growth arrest-specific 6
TCCCCGTACT	30		No Match
AAAAGCTTGA	28	349933	ESTs
TACAAGAGGA	28	349961	RP L6
CGACGAGGAG	26	9999	EMP 3
TGGTACACGT	25	279574	GRIM19
ACTTACCTGC	25	174031	cy c oxidase subunit VIb
GGAGGGATCA	25	6196	integrin-linked kinase
CCTCAGGATA	25		No Match
CCCAGAGACC	23	21223	calponin 1, basic, SM
TCAGACAAAA	23	66881	dynein
CGCCGCGGTG	23	4835	TIF 3, subunit 8 (110kD)
TCCGGCCGCG	23	171774	hypothetical protein
GTGCTGGAGA	23	53125	sn RnP D2 (16.5kD)
AAGAACCTGT	22	75617	collagen, type IV, α2
GCGACCGTCA	21	273415	aldolase A
ACAACTCAAT	21	75922	brain protein I3
TCTGCCTATG	21	90291	laminin, beta 2
GTTAACGTCC	21	178391	ribosomal protein L44
AAAAAGCAGA	21	75428	SOD 1, soluble
AGAAAGATGT	19	78225	annexin A1
TGATAATTCA	19	171625	CDS MGC14697
CAGGCCCCAC	19	256290	S100 protein A11
GCTTGGATCT	19		No Match
ACGTTCTCTT	19		No Match
GAAACCGAGG	18	279813	hypothetical protein

SUMMARY Top 1000	
Total	1000
Reliable Matches	858
Unreliable Matches	142

TABLE 5.3. NON-REDUNDANT NHMC TRANSCRIPTOME (TOP50).

This transcriptome represent the most stringent application of the SAGE criteria but at the expense of a large amount of data. The increase in appearance of No Match entries and the lack of high abundance genes reflects the removal of gene redundancies. When the unreliable tag matches are removed from the 2° transcriptome the top 1000 tags are sampled from the revised or nr transcriptome.

which also indicates the breakdown of mapping with regard to reliable and complete mapping.

Transcriptome	Matches	mRNA	3'EST	No Reliable Match
Primary	83.2%	-	-	16.8%
Secondary	95%	79%	16%	5%
Non redundant	85.8%	58.2%	27.6%	14.2%

TABLE 5.4. SUMMARY OF THE COMPLETE MAPPING OF TAGS TO GENES.

Tags were mapped to genes using the SAGEmap databases. Because of the tag redundancy in the primary transcriptome a breakdown of mRNA and EST mapping was considered uninformative. While the non-redundant transcriptome offered the most reliable mapping a high degree of mapping was to uncharacterised transcripts or failed to match. The secondary transcriptome offered a data set between the two where tag redundancy was removed but the mapping would be sensitive to alternative transcripts.

The nr transcriptome would be expected to be the most accurate data and thus provide the highest degree of confidence in mapping. However, the nr transcriptome had a high level of no-match tags. The peculiar nature of these data derives from the sample used to work out the level of matches and non-matches. When redundancies are removed from the sample, (top 1000 tags) more data take their place. So, while there is more reliable mapping in the 'nr' dataset, the inclusion of lower abundance genes increases the level of 'No Match'. This explains why the 2°transcriptome has the lowest level of no match (multiple UniGene entries are allowed but the tag redundancies have been removed).

5.3 VALIDATION OF SAGE LIBRARY AS A CATALOGUE OF TRANSCRIPTION

Experiments were designed that would determine if there was a correlation between the levels of a gene in the SAGE library and the levels determined using traditional methods. In order to determine whether the mapping information generated from this SAGE analysis was a reliable map of gene transcription, selections of genes were used as probes for hybridisation, PCR experiments and *in silico* mining. Use of three methods for verification provides alternative ways of assessing the integrity of the NHMC SAGE library. Firstly, hybridisation acknowledges the presence or absence of a gene in a conventional and technically straightforward fashion. The use of PCR, in particular quantitative PCR, facilitates the comparison of relative levels of gene expression, generally more specific than hybridisation experiments. Finally, examining

the expression levels of the same tags in other SAGE libraries compares expression levels of ubiquitously expressed genes such as actins, ribosomal proteins and metabolic enzymes.

NHMC n	Gene	Accession	UniGene ID	Seq. Verified	UniGene ID (revised)	[DNA] nM	Size kb
505	TAGLN	AI264733	Hs.75777	VERIFIED		124	1.0
445	α2-SMA	AA565535	Hs.195851	VERIFIED		115	1.3
389	SPARC	AI805515	Hs.111779	ESTs	No Cluster	35	3.1
239	FN-1	R62612	Hs.287820	VERIFIED		141	1.0
147	CCND1	AA133137	Hs.82932	VERIFIED		38	3.0
147	CCND1	AA599485	Hs.82932	VERIFIED		35	2.9
135	CD151	AI682067	Hs.75564	VERIFIED		84	1.7
74	(GLA)	AA983329	Hs.75742	COLL IV a2	Hs.75617	158	0.87
73	HSP70	AA173766	Hs.75410	VERIFIED		56	2.8
37	MFGE8	AA001073	Hs.3745	VERIFIED		53	2.3
33	CRYAB	AA613030	Hs.1940	VERIFIED		151	0.93
27	VIM	AA487812	Hs.2064	VERIFIED		99	1.6
24	ENIGMA	AI566597	Hs.102948	VERIFIED		59	2.5
23	P4HB	AA128412	Hs.387107	VERIFIED		138	1.3
22	DAXX	H49831	Hs.180224	VERIFIED		152	0.86
17	QSCN6	R84988	Hs.77266	VERIFIED		51	2.4
16	LRP1	AI273077	Hs.475478	VERIFIED		14	5.5
14	GRN	AA653437	Hs.180577	VERIFIED		58	2.5
8	SERPIN b6	AA581424	Hs.41072	VERIFIED		22	5.5
6	FK605bp9	AI073739	Hs.8762	VERIFIED		87	1.8
6	TGF-β2	AI292328	Hs.169300	VERIFIED		115	1.1
6	(CTGF)	AA599520	Hs.75511	BASP-1	Hs.79516	36	2.8
6	(DBI)	AA031339	Hs.78888	EPLIN	Hs.10706	167	0.9
5	DCIP-1	AI219163	Hs.36794	VERIFIED		93	1.9
5	MCT-1	AA115687	Hs.102696	VERIFIED		136	1.3
4	LTPB4	AI634901	Hs.85087	VERIFIED		60	2.0
4	TGF-β1	AI433146	Hs.1103	VERIFIED		41	2.3
3	TRAIL-rc2	AA453916	Hs.51233	VERIFIED		135	1.2
3	(ECM1)	AI473375	Hs.81071	TC21	Hs.206097	133	0.91
1	(CSPG4)	AI423029	Hs.9004	ESTs	Hs.260238	188	0.75
0	(ECM1)	AI338404	Hs.81071	BRD3	Hs.86896	100	1.4
0	(X11-like)	R55870	Hs.26468	MAPK P'ase	Hs.20281	120	1.0
0	BMP-5	AI239557	Hs.1104	VERIFIED		120	1.5

TABLE 5.5. PROBES USED TO VALIDATE GENE TRANSCRIPTION IN NHMC.

Each gene fragment was amplified from the IMAGE clone and sequenced to confirm identity. The PCR products were diluted to equimolar concentrations and fixed to nylon membranes then hybridised to complex single stranded cDNA probes as described in CHAPTER 2. Note that not all the requested clones were confirmed as their cluster identified and indicates an uncertainty in the UniGene clustering and IMAGE consortium.

5.3.1 DOT BLOTS DEMONSTRATE THE PRESENCE OF TRANSCRIPTS

A selection of genes was chosen based on a broad level of abundance, as determined by the SAGE library. Each of these genes was obtained as an IMAGE clone and sequence verified (see TABLE 5.5). Fragments of gene sequences were immobilised onto nylon membranes and hybridised using complex cDNA probes labelled with α-^{32}P-dCTP (see CHAPTER 2.4.1). An example of a hybridised filter is shown in FIGURE 5.2.

Individual signals are tabulated against the SAGE derived abundance in TABLE 5.6. A simple visual assessment is in general agreement with the SAGE derived mRNA frequency. The correlation coefficient, r^2, of 0.7 confirms this assessment. Outliner genes, CCND, MYC-B, VIM and CD151 require clarification regarding true abundance, cross hybridisation and tag mapping accuracy.

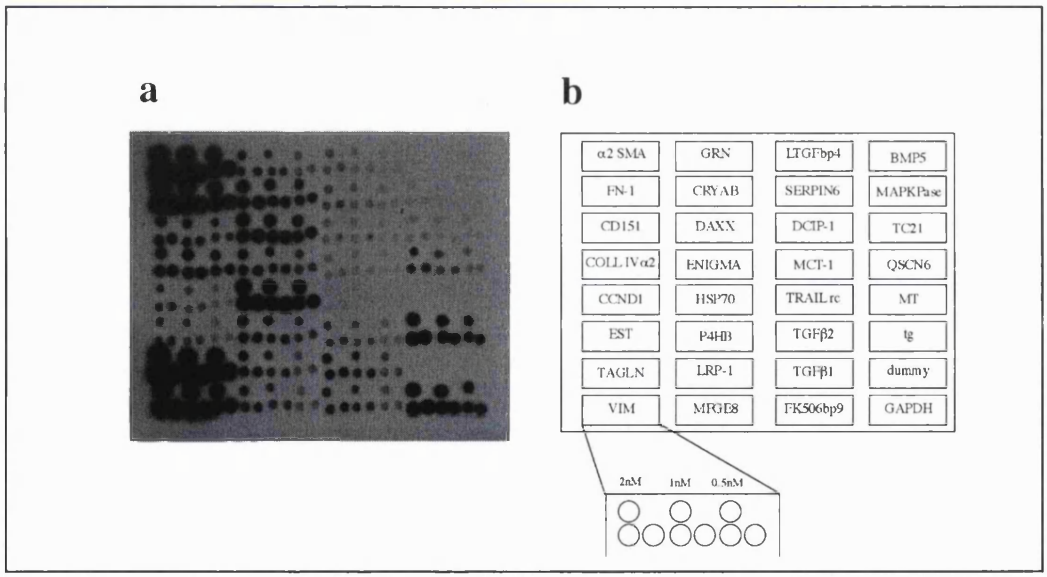


FIGURE 5.2 A&B. A MANUALLY CONSTRUCTED DOT BLOT HYBRIDISED TO LABELLED SSDNA (A).

For construction and hybridisation see CHAPTER2. The data appears in general agreement insomuch as a hybridisation signal is present in the great majority of tags. This hybridisation was repeated with RNA from two other culture series giving a total of 3 independent RNA sources hybridised to the same filters. The figure key (b) lists the genes and targets used. MT (-ve no target), tg (+ve total genomic target), dummy (-ve non-homologous target). Each gene cluster was composed of triplicate spots of 2, 1 and 0.5nM.

Gene	Tag Sequence	IMAGE EST	SAGE n	Hybridisation Signal
TAGLN	ACAGGCTACG	AI264733	252	
α-SMA	AAGATCAAGA	AA565535	207	
FN-1	ATCTTGTTAC	R62612	116	
GAPDH	TACCATCAAT	BC004319	76	
CD151	TGCCTCTGCG	AI859372	70	
HSP70	TGCATCTGGT	AA173766	41	
VIM	TCCAAATCGA	AA487812	18	
CRYAB	GTTTCATCTC	AA613030	17	
ENIGMA	TGTGAGCCCC	AA171520	15	
MFGE8	GGTTGGCAGG	AA001073	15	
COLL IVα2	GTTTATGGAT	AA983329	13	
LRP1	CTCAACCCCC	AI273077	12	
CCND	AAAGTCTAGA	AA599485	12	
MLC-B	CAACTTAGTT	AA071085	11	
P4HB	CCTGGAAGAG	AI469235	10	
GRN	GGAGGTGGGG	AA653437	9	
SERPIN B6	ATGATGCGGT	AA581424	7	
QSCN6	CTTGATTCCC	AA947615	5	
MCT-1	AAGATAATGC	AA115687	5	
LTGFbp4	CCCTCTCCCT	AA632997	4	
TRAILrc	ACCAAATTAA	AA453916	1	
FK605bp9	GAATAAATGT	AI073739	1	
TGF-β1	GGGGCTGTAT	AI433146	1	
TC21	AAGTTTATAG	AI473375	0	
BMP-5	TGGCTGCCAC	AI239557	0	
MAPK P'ase	CAGCTACCCG	R55870	0	
EST		AI423029	0	
TGF-β2	TTATGTATCA	AI292328	0	
DCIP01	ATTTCTCATT	AI219163	0	
DUMMY	-	-	-	
BLANK	-	-	-	
Total G	-	-	-	

TABLE 5.6. INDIVIDUAL HYBRIDISATION SIGNAL COMPARED TO SAGE DERIVED FREQUENCY.

Although a crude measure of abundance it appears that as the SAGE derived frequency increases so does the hybridisation signal (r^2=0.7). Outliner genes, CCND, MYC-B, VIM and CD151 require clarification regarding true abundance, cross hybridisation and tag mapping accuracy.

The hybridisation signals require two normalisation factors before they can become semi-quantitative. First must be considered the relationship between the target (immobilised gene fragments) and the query (labelled ssDNA). The target must saturate the query, otherwise a non-linear relationship will result, where increasing levels of query will not increase hybridisation signal. This is not generally an issue with low and medium abundance genes but can present problems with high abundance genes such as

TAGLN and VIM, as well as those used to normalise filters between experiments (i.e. housekeeping genes such as GAPDH and Actin). This saturation of target by the query will lead to inaccurate normalisation of experiments. The second normalisation factor requires the calculation of molar relationships between the targets. The chemical properties of nucleic acids mean the same weight (μg) of two species of DNA that differ in length will have different molar values.

The molar concentration of gene fragments was determined from the length of PCR products and a molar conversion relationship (EQUATION 2.1). From this relationship, it can be seen that, without the molar value for each target, the hybridisation signals will need to be adjusted accordingly. This issue was partially addressed by immobilising equimolar amounts of target in the construction of the blots. Even considering this normalisation, the nature of labelling by reverse transcriptase means that the efficiency of activity, the length of probe and relative %CG will require acknowledgment for absolute quantitation. Relative semi-quantitative measurements were all that was required.

Cross reactivity within gene families is also a contributing factor in hybridisation experiments and can create inaccurately high signals. Although these errors will generally be constant across experiments, with corresponding target they can prove problematic when trying to accurately quantify individual signals relative to a number of genes in the same experiment, and are insensitive to transcript variation.

Nevertheless, qualitative and semi-quantitative measures were seen and the correlation of NHMC tag frequency and hybridisation signal was determined (r^2= 0.7), indicative of a high similarity. Uncertainty can arise for two reasons. First, given that gene families will cross react to a certain degree in hybridisations, and so bias the hybridisation signals to higher measured values, the hybridisation experiments will bias toward an over estimation of the true gene abundance. Secondly, different transcripts from the same gene may not be represented by the same SAGE tag, thus leading to an underestimation of true abundance when compared to the hybridisation signal. Combined, these phenomena lead to over-estimation of hybridisation signal and an under-estimation of SAGE determined abundance, both of which constitute a large source of potential error.

5.3.2 REAL-TIME RT-PCR QUANTIFIES ABUNDANCE OF TRANSCRIPTS

Real time RT-PCR (rtRT-PCR) is a powerful technique that is increasing in use. The ability to track a PCR reaction with accuracy means that obtaining quantitative data from PCR reactions no longer requires large investment in optimisation. A more detailed description of rtRT-PCR can be found in METHODS 2.9. A selection of genes from the NHMC transcriptome was used to design primers for use in quantitative PCR reactions (see TABLE 5.7). They were chosen to represent a broad range of abundances, as determined in the NHMC SAGE library. Experiments were conducted in duplicate, using RNA from three separate culture series, giving a total of six experimental data points for each gene.

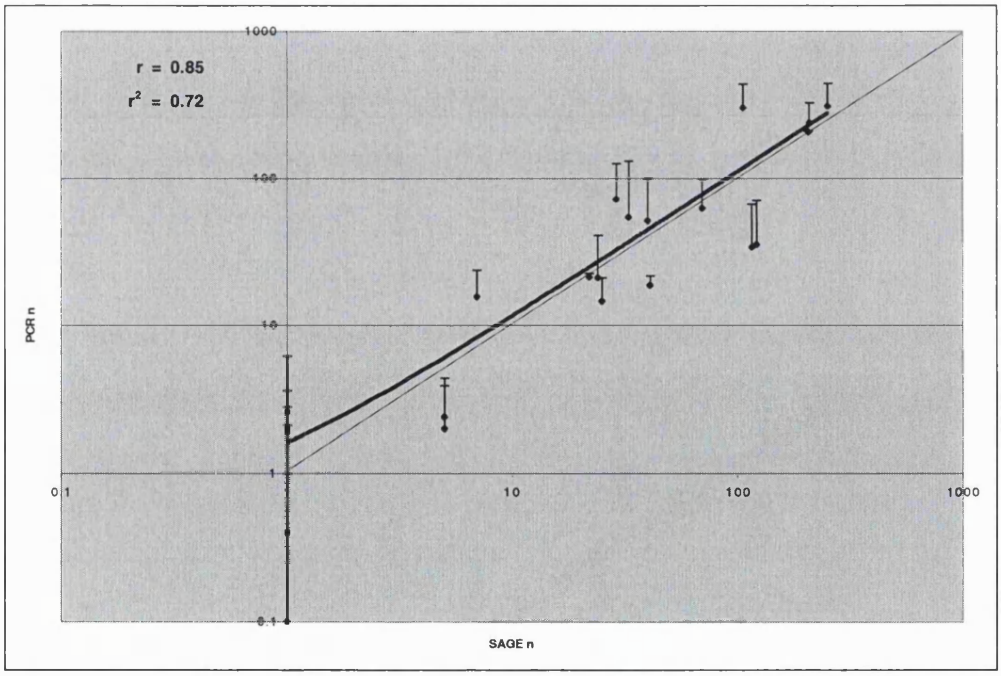


FIGURE 5.3. COMPARISON OF SAGE AND RT-PCR DETERMINED ABUNDANCE.

Graphical representation of quantitative data obtained from SAGE analysis (x axis) and PCR determined abundance (y axis). The 'line of similarity is shown as a 45° diagonal and the bold line is the trendline. The Pearson coefficient 'r' indicates a high degree of similarity as does the r^2 correlation value. Only positive error bars are included, as negative values cannot be plotted accurately on a log scale

Relative quantitative values for each amplicon were determined from a standard curve, generated from serial dilution of a target sequence amplified in the same experiment. Data was imported into spreadsheets, average levels were calculated and, as a normal distribution, the standard deviation from the mean was used to estimate

error. A graphical representation of the two datasets is shown in FIGURE 5.3. The 'line of similarity', 45° to the axes, shows that the data from each experimental methods increases together with a correlation value of $r^2 = 0.72$ indicating again a high degree of similarity.

GENE	UniGene ID	PRIMERS 5'-3'		Amplicon (bp)
		Forward	Reverse	
CSPG	Hs.9004	CATCCTGCCCCTGCTCTTCTACCT	CCGGACCCCTGGGACTATCTC	418
CTGF	Hs.75511	ACGGCGAGGTCATGAAGAAGAACA	TGGGGCTACAGGCAGGTCAGTG	523
TAGLN	Hs.75777	AAGAAAGCGCAGGAGCATAAGAGG	AGCCCAGGGAGGAGACAGTAGAGG	409
SPARC	Hs.111779	TGGGCAAAGGGAAGTAACAGACAC	CAACCGATTCACCAACTCCACTTT	446
IGFbp7	Hs.119206	CGTGAAGAGCCGCAAGAGG	AGATACCAGCACCCAGCCAGTTAC	448
HSP70	Hs.75410	AAGAGGGAGGGGGAGAAGAACATC	GAGTCGAGCCACCAACAAGAACAA	463
GLA	Hs.75742	TGACCTGCAGGACGAAACC	TGCTACAGGGGGATACAAAAT	400
GAL1	Hs.227751	AATCATGGCTTGTGGTCTGG	CATGGGCTGGCTGATTTC	428
COLLIVa1	Hs.119129	CCTGCCGGGCCTACTGGT	GGCACGGTGGGATCTGAATGGT	450
COLLIIIa1	Hs.119571	GGGGTCCTACTGGTCCTATTGG	CCTGGGGGTCCTGGGTTAC	400
COLLIVa2	Hs.75617	CTGGCCCTGGAAAAGATG	CAGGTTGGGCCCGGATACAGGTT	478
CD151	Hs.75564	TGGCGGGCACTGTCGTCAT	GCAGCCGCCCTCCACCTTGTAGAT	428
ACTA	Hs.727	TCGGGGAGAACGGGTATGTGGAGA	CAACCGCTGGATGCTGCTGGAGA	407
FBN1	Hs.750	TCCCGGATTTACCCAACACCATAC	GCCTGCGCAGAGCCACATT	435
TGFβ1	Hs.1103	CTGCCTCCTCCTGCCTGTCT	CCCGCCTGGCCTGAACTACT	309
TGFβ2	Hs.169300	CTGCGTGTCCCAAGATTTAGAAC	TCAAGTGAGGCGCGGGATAG	514
GAPDH	Hs.169476	GCGGGGCTCTCCAGAACATCATCC	GCCAGCCCCAGCGTCAAAGGTG	470
ASMA	Hs.195851	AAGGCCAACCGGGAGAAAATGACT	CCGATGAAGGATGGCTGGAACA	467
βACT	Hs.180952	ATGATATCGCCGCGCTCGTCGTC	AGGTCCCGGCCAGCCAGGTCCAG	547
SERPIN B6	Hs.41072	AATCATGCTTCCGGACGAGACCACT	AAGCGGCCGCAGAAGAGAATCC	425
PGM-1	Hs.1869	GGCCACCACCCTGACCCCAACC	GCGCCCTCAGCTTCCACCTCCTC	539
DCIP	Hs.3674	AGAGCCCTGAGAACAATGACCTTA	TCTGTGCCACCTGATCCTTCTTC	396
FK506bp9	Hs.302749	CAGCCTCCGCCTCCCGTATTCA	GCCCAGCCCCCAGCCCTATTTAT	317
MCT-1	Hs.102696	GCGCGGCTGGCTCTCGT	TGGTGTGGCAGGATAAAAGGATA	435

TABLE 5.7. GENES USED TO CREATE TEMPLATES FOR RTRT-PCR AND THE PRIMER SEQUENCES USED.

Characterised cDNA was used to design the primers and where possible the amplicon was biased toward to 3' end of the transcript. In keeping with instructions from the manufacturer, amplicons below 500bp were preferred. For further information, see CHAPTER 2 9.

Thus there appears a high degree of similarity between data derived experimentally from SAGE and that from more conventional hybridisation and PCR experiments. This is turn provides strong evidence that SAGE is an accurate and reliable catalogue of genes transcribed in NHMCs. As can be seen in FIGURE 5.3 a more accurate quantitation is obtained from high abundance genes. Graphically this is illustrated as a small distance from the line of similarity for high abundance genes and the larger variation from this line and curving of the trendline for lower abundance genes. These properties are consistent with a lower reliability of the SAGE data with lower tag counts.

5.4 SAGE TAGS ARE PRESENT IN OTHER LIBRARIES AT SIMILAR RELATIVE LEVELS

Together with gene identification, SAGEmap allows public access to all registered SAGE libraries (some 200 to date). This information is useful in many ways and was applied here to determine the frequency of a selection of tags several SAGE libraries (www.ncbi.nlm.nih.gov/sagemap). Conventional Northern blots that assess gene expression across a variety of time scales, tissues or cell types are commercially available, but require the application of experimental samples and dynamic monitoring. One of the attractions of SAGE data is the simplicity of data archiving, but the digital nature of the data also allows the *in silico* analysis of gene expression across experimental samples. When applied to the question 'how a gene is expressed in a variety of tissues or cells' one need only look for the tag or tags that correspond to that particular gene in the SAGEmap database and its presence across all libraries can be determined. This was used in a further analysis of the NHMC transcriptome (CHAPTER 6), but here, housekeeping genes in SAGEmap were mined for their respective frequency of appearance in a variety of libraries. The results demonstrate that all the tags are indeed present in almost all of the libraries at comparable levels which suggests that the SAGE data in the NHMC library is comparable to other SAGE libraries (see TABLE 5.8.)

Tag Sequence	293	Cere	Fibro	H216	Heart	Kidney	Liver	NHA	NHMC	PR317	Prostate	UniGene ID	Gene
AAGATCAAGA					26			14	110	31		1288	ASMA
CACAAACGGT	39	10	31	63	13	7	7	9	59	37	8	195453	RPs27
TACCATCAAT	34	13	9	4	43	4	17	54	35	27	2	169476	GAPDH
AGCACCTCCA	23	13	20	36	15	8	11	24	32	30	7	75309	EF2
TGCATCTGGT			10	7	5		13	4	18	3	4	75410	HSP70
GTGAAACCCC	30	112	28	55	8	43	8	29	28	5	47	44396	Coronin
TTGGGGTTTC	4	4	34	4	38	47	17	43	18	31	17	62954	Ferritin H
GTTGTGGTTA	2		25	6	23	9	30	4	26	45		75415	b2 Microglc
GGCAAGCCCC	10	3	23	27	11	8	7	5	13	23	32	76067	HSP27
AGCCCTACAA	4	54	5	21	107	88	21	16	15	60	36	95243	TEF-A
CTGGGTTAAT	46	4	52	13	15	3	14	14	25	47	77	101047	TEF-3
AATCCTGTGG	27		8	25	6	5	3	6	12	35		111334	Ferritin L
ACTTACCTGC	5	6		3	18	4	6	3	6	5	2	174031	CyC Ox Ivb
GAAGCAGGAC	6	7	43	7	5	3	3	25	11	7	7	180370	Cofilin
TGTGTTGAGA	82	15	30	72	7	22	9	40	38	68	14	181165	EF1a
CGCTGGTTCC	15	9	22	26	8	7	10	12	14	14	9	181165	EF1a
GTTCGTGCCA	9	4	15	23	9	3	3	4	17	8	10	195464	Fillamin

KEY

Tag Freq

0
1 to 10
11 to 40
41 to 70
>70

TABLE 5.8. DIGITAL NORTHERN OF A SELECTION OF HOUSEKEEPING GENES. The various libraries are described in TABLE 4.3a. Although there are wide variances between the frequencies of tags in each library the presence, at this level, in practically all libraries indicates a constitutively expressed gene. Note that α2SMA is present predominantly in astrocytes (NHA), mesangial cells and smooth muscle cells (PR317). The grey scale has been used to assist in visualisation and is based on frequency distribution of the respective tag. The higher the frequency the darker the shade (see key)

5.5 SUMMARY OF SAGE VALIDATION

The use of three independent methods of validation has demonstrated that the SAGE data in NHMC library is valid and reliable. In the first method, RNA from three independent culture series was labelled with α-^{32}P-dCTP and hybridised to the immobilised gene fragments. The presence of a signal above the background demonstrated the presence of the mRNA. As well as qualitative hybridisation data, a semi-quantitative relationship between the SAGE and hybridisation data was demonstrated where signal intensity increased with the SAGE derived abundance. A more accurate measure of abundance was determined using real time RT-PCR. The data from these experiments again demonstrated a high correlation with the SAGE. Finally tags from housekeeping genes were used to mine *in silico* SAGEmap data from other SAGE libraries. This added evidence for the validity of the NHMC library with many housekeeping genes predicted to be expressed across all cell types represented by tags across all libraries.

5.6 DISCUSSION

When designing a SAGE analysis it is important to consider the size and complexity of the transcriptome under investigation. A pure cell culture will have advantages over a heterogeneous population of cells or tissue where more than one cellular transcriptome is being considered (pure cell vs. micro environment), and a smaller genome will significantly reduce required tag collection and thus errors. Clearly the size of the genome will be dependent upon the organism under investigation; a SAGE library in yeast *sp.* will require a lower sampling level than any mammalian system. But in complex organisms, the histology of the target for which one intends to analyse is also of great importance. In the context of this study, this issue is particularly important. The mesangial cell accounts for approximately 30% of glomerular cells and proximal tubule cells are some 60% of kidney mass (Mene et al., '96, Virlon et al., '99). Generating a transcriptome from isolated glomeruli or bulk kidney tissue will result in a library composed of a variety of cell transcriptomes and thus a more complex picture of transcription. This will be further discussed in CHAPTER 8, but must be appreciated that when sampling populations, increasing the number of populations to be sampled exponentially increases the level of sampling required.

From the SAGE library generated in CHAPTER 4, a transcriptome of the NHMC was constructed using mapping information currently available at the NCBI. This mapping information provides a list of all the genes that map to the tags in this SAGE library and information regarding their respective UniGene clusters. Transcriptomes were generated that contained mapping data at three levels of reliability based on the redundancy of the tags and genes. The primary transcriptome provided raw mapping data for all tags and showed striking redundancy where a unique tag mapped to more than one gene. The secondary transcriptome, generated by removing tag redundancies, contained redundancies where UniGene clusters would map to more than one tag. While possibly also inaccurate, this transcriptome theoretically contains information on the alternative processing of transcripts which will be addressed in the next chapter (CHAPTER 6). The final 'non-redundant' transcriptome was the most accurate transcriptome, but contained far fewer mapped tags. While the mapping data was more accurate in the 2° and nr transcriptomes, the inclusion of lower abundance tags led to fewer assignments.

A selection of genes, representing several abundance levels from the 2° transcriptome was used to verify the reliability of the NHMC-SAGE library. These genes were used to construct dot blots that were hybridised to complex cDNA probes. The hybridisation signals demonstrated a high degree of correlation between SAGE determined frequency and hybridisation signal. In addition, PCR primers were designed using reference sequences and used in real-time RT-PCR experiments, generally more quantitative that hybridisation experiments yet also showed a high correlation.

All the above experiments contained a number of outliers. In order for these to be rationalised in an experimental setting, the differences and similarities between the techniques need to be assessed as a function of their ability to delineate between actual and experimental abundance of a gene. The 'actual' abundance of a gene is a general term that will be defined as the cluster of transcriptional products from a gene. Thus, the actual abundance reflects the clustering of ESTs to genes and will include of all products from that gene. This definition will create the first inaccuracy of quantitation for all techniques. Hybridisation will favour the detection of all gene products, SAGE will be sensitive to different transcriptional products, while PCR will be dependent on the placement of the primers. Therefore, hybridisation will be more accurate for actual transcriptional abundance, while SAGE and PCR will be more accurate for transcriptional variation. Other properties of gene expression may also add to the differences seen within the techniques. Cross hybridisation between highly similar genes or homologous regions will inflate the hybridisation signal, when compared to the SAGE or PCR determined abundance. Inhibition of PCR reactions will result in under representation and variation between experiments. As all the techniques require the initial generation of cDNA, any mis-priming of the first strand oligo-dT will result in under representation in SAGE and PCR but should not affect hybridisation signals. Finally, a SAGE analysis has inherent technical failures like sequencing and mapping errors that can simultaneously inflate and deflate gene abundance (discussed in CHAPTER 4).

Even with these considerations, these experiments clearly demonstrated a strong correlation between abundance determined by SAGE and that determined by rt RT-PCR and hybridisation. From these experiments, it would appear that the NHMC SAGE library was a reliable indicator of transcription, and that the presence of a gene in the

secondary or non-redundant transcriptomes was real, but that a significant portion of genes may be masked in the 1° transcriptome.

A significant number of tags (4,110) failed to map to the reliable databases. This appears a common feature in many SAGE analyses and poses three interesting questions. First, the tag may be a by-product of the technical protocol and so represent nothing, i.e. the tag is an error. In this case, the abundance of a gene will be artificially increased while another gene will be incorrectly reduced. This will be of little significance in genes for which there is a large amount of data, but genes of low abundance will be compromised. Counting the same ditag once protects from the inclusion of PCR artefacts and the high levels of some no-match tags present (greater than 4) suggests they are real tags. The second possibility for a 'no-match' tag may be the presence of an un-described gene. As HGMP describes nearly all the genes in the human genome, this seems unlikely.

Tag Sequence	LG	HG	SUM	Reliable	Un-reliable mapping
CAAGCATCCC	16	9	25	-	Hs.151242 (SERPIN G1) Hs.327884 (EST)
CCTCAGGATA	9	14	23	Hs.170009 (TGFα)	Hs.184601 (SLC7A5) Hs.252723 (RP L19) Hs.335919 (EST) 148Unclustered ESTs
ACGTTCTCTT	8	10	18	-	No Match
GGGAAGCAGA	12	5	17	-	Hs.40500 (RER1 homologue) Hs.78713 (SLC25A3) Hs.12284 (JAM1) 8 Unclustered (ESTs)
GTTGGGTTAA	11	4	15	-	No Match
ATACAAGAGC	1	10	11	Hs.727 (Inhibitin,βA)	2 Unclustered (ESTs)
GTGCTGAAGG	9	1	10	-	Hs.294088 (GAJ) 1 Unclustered EST
AGGCTCGGAA	6	4	10	-	Hs.119122 (RP L13a)
CTGCGAGTGA	6	4	10	Hs. 234680 (fer-1-like3)	
CCAGGGCAAC	4	5	9	Hs.252923 (EST)	Hs.151413 (GMF β)

TABLE 5.9. TOP TEN TAGS THAT INITIALLY FAILED TO MATCH IN THE RELIABLE DATABASE.

Subsequent queries revealed reliable mapping for a further 4 tags Frequency in low glucose (LG) and high glucose (HG) is indicated and where there was a match in the total mapping database the EST is described. The high frequency of the tags suggests un-annotated gene or transcript.

The final, and most likely reason for un-matched tags, would be an un-described transcript. The clustering of ESTs to genes, as in UniGene, is a dynamic process and will not be complete until all transcripts for each gene are described. It may be that an un-matched tag is the result of transcript specific to the system (possibly indicative of

phenotype) and that this transcript has not been reliably cloned and characterised. When the un-matched tags are queried in the 'total mapping' database, many of them map to ESTs. This mapping is not included in the 'reliable' mapping database as the annotation of sense to the sequence could not be determined, i.e. the real 3' end of the gene was unclear and so reliable tag assignment could not be made (see TABLE 5.8). That said, even with the total mapping, some tags remained unassigned. These un-matched tags require clarification and eventual assignment.

The digital nature of SAGE facilitates not only the storage and retrieval of data but also the comparison between different experiments. With more conventional transcription analysis, complex normalising matrices are required for each set of experiments to allow comparison. This is partly due to the experimental protocol and introduces errors across platforms. Yet, even normalisation, where it exists, provides relatively low-resolution comparisons with large investment in laboratory resources. SAGE libraries derived from epithelial cells can be compared directly to SAGE libraries derived from astrocytes at a resolution much higher than hybridisation experiments can currently achieve. These comparisons, or 'digital tissue blots', can be constructed with relative ease, requiring basic computing power and no 'wet' laboratory time. The digital blots can provide insight to the restricted expression of genes across tissues and can facilitate the identification of genes whose expression is restricted to phenotype whether that phenotype is between related cells or cells obtained *in vivo* or cultured *in vitro*. This will be investigated further in CHAPTER 6.

In summary, a SAGE library of NHMCs was sampled and 43,358 tags were mapped to 20,382 genes, which constitutes a transcriptional profile of cultured normal mesangial cells and a resource for future investigation. This accuracy of gene to tag assignment was verified by standard hybridisation techniques and the relative abundance of a selection of genes was determined by real-time RT-PCR. All experiments conferred a high degree of reliability to the SAGE library and so this NHMC library can be considered a transcriptome of NHMC. To our knowledge this is the first SAGE library created from cultured normal human mesangial cells and provides useful information of the transcription characteristics of the mesangial cell. This, in turn, will provide information defining the functional characteristics of the mesangial cell and contribute to the understanding of its role in renal physiology and pathogenesis.

CHAPTER 6

6 DISCRIPTION OF GENES WITHIN THE NHMC TRANSCRIPTOME

6.1 INTRODUCTION

Experiments in CHAPTER 4 demonstrated that a SAGE transcriptome can be used to classify cell types, those derived from similar lineages had similar transcriptomes than unrelated cells. Experiments in CHAPTER 5 demonstrated that the frequency distribution of sampled SAGE tags was consistent with the abundance of genes in the NHMC transcriptome. In this chapter, the transcriptome of cultured NHMC is described and digital northerns are generated to identify genes that appear restricted to NHMC or related cells.

From the analysis in CHAPTER 5 the SAGE determined transcriptome exists at three levels. An analysis of the primary transcriptome can be confusing and much time can be spent attempting to distinguish between genes that are represented by the same tag. The problem of tag redundancy can be alleviated if subsets of redundant tags are mapped to an 11^{th} base as has been suggested in many studies (Velculescu et al., '95, Velculescu et al., '97). Mapping the 11^{th} base will be discussed in CHAPTER 7. On the other hand, the most reliable map or the non-redundant transcriptome falls short of a complete transcriptome comprising of only 8,774 entries. There is a high likelihood that important data, such as alternative transcription, is lost in the stringent application of the SAGE criteria.

The middle ground between the 1° and 'nr' transcriptomes is the secondary transcriptome where the tag redundancy is removed but multiple tags are permitted to map to the same gene. This database will contain both the unique non-redundant mapping together with the multiple gene representation. For this reason the secondary transcriptome was used to classify the data into functional, structural and other groups.

This chapter contains a description of the types of genes present in the highest abundance class overall and a classification of genes based on function. While gene ontology is particularly subjective, such large amounts of transcription data require some sort of grouping. A more complete catalogue of genes is presented in APPENDIX 4. In the case where the presence of a particular gene is significant but the abundance falls below the threshold of the table guides (generally top 20 genes overall, TABLE 6.1 and top 10 genes in each group, TABLES 6.2 - 6.8), then the relative abundance and

frequency is described in a final section concerning functional significance in DN (TABLE 6.9)

6.2 THE NHMC 2° TRANSCRIPTOME

6.2.1 HIGH ABUNDANCE GENES

The most prominent genes in the library were those concerned with protein synthesis (ribosomal proteins), matrix remodelling and the cytoskeleton (TABLE 6.1). Of particular abundance are the genes for transgelin (TAGLN, formally SM22alpha) and osteonectin (SPARC). Transgelin was originally believed to be specific to smooth muscle cells, but is highly represented in most normal mesenchymal cells (Lawson et al., '97).

Tag Sequence	Total Freq.	UniGene ID (Hs.)	Gene Description
ACAGGCTACG	562	75777	TAGLN (transgelin)
AAGATCAAGA	445	1288 14376 195851	Actin (3 isoforms)
ATGTGAAGAG	424	111779	SPARC (osteonectin)
CACAAACGGT	266	195453	RP S27 (metallopanstimulin 1)
TTTGCACCTT	260	75511	Connective tissue growth factor
ATCTTGTTAC	255	287820	Fibronectin 1
GACCGCAGGA	248	119129	Collagen, type IV, alpha 1
CATATCATTA	232	119206	Insulin-like growth factor binding protein 7
GGAGTGTGCT	232	9615	Myosin regulatory lc 2, smooth muscle isoform
AAGACAGTGG	199	296290	Ribosomal protein L37a
TTGGTGAAGG	192	75968	Thyroxin, beta 4, X chromosome
CCCATCGTCC	182	Mito	Tag matches mitochondrial sequence
GCCCCCAATA	181	227751	Lectin, galactoside-binding, soluble, 1 (galectin 1)
TACCATCAAT	157	169476	GAPDH
TGCCTCTGCG	151	75564	CD151 antigen
AGGCTACGGA	149	119122	RP L13a
AGCACCTCCA	143	75309	Eukaryotic translation elongation factor 2
GACCAGGCCC	138	300772	Tropomyosin 2 (beta)
ATTCTCCAGT	136	234518	Ribosomal protein L23
CCGTGACTCT	131	296267	Follistatin-like 1
GCATAATAGG	125	184108	Ribosomal protein L21

TABLE 6.1. TOP 20 TAGS AND CORRESPONDING GENES EXTRACTED FROM THE 2° TRANSCRIPTOME. The levels from both low and high glucose sets are shown together with the UniGene Hs. Cluster ID.

TAGLN is found in high abundance in SAGE libraries of normal cultured astrocytes and micro-dissected prostate tissue (histologically rich in smooth muscle cells). In this library the TAGLN tag was the most abundant tag, present at 1.2%.

TAGLN is reported to be transformation sensitive, thus its expression was investigated in the commonly used SV40 immortalized human mesangial cell lines where it was demonstrated to be strongly down regulated (see CHAPTER 7). Transgelin shows structural similarity to the myofibrillar regulatory protein calponin that is also strongly down regulated upon transformation (Shapland et al., '88). TAGLN binds actin (Kd $7.5x10^5$ M^{-1}) and converts loose actin filaments to tangled cross-linked mesh (Shapland et al., '93).

The matricellular protein SPARC is also highly represented in many libraries and was present at high abundance in this library. SPARC is a secreted PDGF binding glycoprotein involved in cell-ECM interactions and the cell cycle (Pichler et al., '96, Bradshaw et al., '99). SPARC has also been associated with positive regulation of Type I collagen and the pro-fibrotic cytokine TGF-β1 (Francki et al., '99, Bassuk et al., '00). SPARC is highly represented in many SAGE and EST libraries.

As expected, actin was highly represented (1%) though due to 3' similarity the tag cannot distinguish between α 1 (skeletal muscle actin), α 2 (smooth muscle actin) and γ 1 (non-muscle actin). Taking into account the mesenchymal origin of NMHC and smooth muscle cells, it is likely that α-2 is the predominant actin. There is strong evidence that the presence of such large amounts of α2-SMA is more indicative of cellular activation, through serum starvation, than an accurate marker of smooth muscle like cells. (Stephenson et al., '98, Kitamura and Ishikawa, '98). As expected actin tags are highly represented across all SAGE libraries. Large amounts of actin and various forms of myosin and tropomyosin are consistent with the contractile functions of the mesangial cell.

Fibronectin (FN-1) is well represented in this and many gene libraries. FN-1 is a high molecular weight glycoprotein present on cell surfaces and in serum. The primary function of fibronectin is to facilitate the interaction of cell membranes with collagenous matrix. Well characterised in mesangial cells, increases in FN-1 production, together with other matrix proteins, is one of the primary events in the expansion of mECM in diabetic nephropathy (Jackle-Meyer et al., '95, Elbein and Kaushal, '99).

Activin A (ACTA, Inhibitin, Beta A) is a member of the TGF β superfamily of pluripotent growth factors and has activities in development, differentiation, and tissue fibrosis (Matzuk et al., '95, Peng and Mukai, '00). NHMC express ACTA, as well as members of the TGFβ family of growth factors, although at relatively low level (e.g. TGFβ1, 0.008%). TGFβ1 is a central growth factor in wound repair and the development of fibrosis and is believed to drive DN.

Mitochondrial sequences are present at a similar level in all SAGE libraries, yet the mapping database does not map these to any of the genes present in the mitochondrial genome. One report described the mapping of tags derived from mitochondrial sequences that represented genes involved in oxidative phosphorylation, which would indicate constitutive expression (Welle et al., '99). This may be important in this model, as oxidative phosphorylation is a target process disrupted in DN and a postulated to be a primary causative mechanism.

Thymosin β4 (TMSB4X) is a member of a multi-gene family of growth and differentiation factors. TMBS4X can inhibit the migration of macrophages and induce differentiation in T cells and is believed to take part in the control of actin polymerisation in non-muscle cells (Clauss et al., '91). The expression of TMBS4X can be detected in many tissues but is particularly high in spleen, thymus, lung, and peritoneal macrophages (Li et al., '96). The gene can exist as a number of transcripts, particularly in lymphocytes, which are believed to have a unique splice variant. The requirement of TMBS4X in metastasis in melanoma cells suggests that cell cycle involvement is also an activity of this protein (Clark et al., '00).

6.2.2 CATEGORIES ACCORDING TO FUNCTION

6.2.2.1 PROMINENT CYTOSKELETON GENES

Genes involved in the cytoskeleton were also present in high abundance in the NHMC transcriptome. In particular, the mesenchymal origin of NHMC is apparent with specific isoforms of myosins and tropomyosins (TABLE 6.2)

Of particular note is the myosin lc2 isoform, which is smooth muscle specific. Other actin-interacting proteins like profilin, cofilin and filamin-Aα are also well

represented. Gesolin is a widely distributed protein that dissolves actin gels in a Ca^{++} dependent mechanism (Kwiatkowski et al., '86). Transgelin2 is a paralogue of the more abundant transgelin whose precise function is unknown but likely to be similar to transgelin (Stanier et al., '98). Crystalline AB, a small heat-shock protein associated with the cytoskeleton, shows high abundance in NHMC. This protein is constitutively expressed in the lens of the eye, myocardial cells, and kidney epithelium (Dubin et al., '90).

Cytoskeleton			
Tag Sequence	Total Freq.	UniGene ID (Hs.)	Gene Description
ACAGGCTACG	562	75777	Transgelin
GGAGTGTGCT	232	9615	Myosin regulatory light chain 2, smooth muscle isoform
TTGGTGAAGG	192	75968	Thymosin, beta 4, X chromosome
GACCAGGCCC	138	300772	Tropomyosin 2 (beta)
GGGGAAATCG	73	76293	Thymosin, beta 10
CGAGGGGCCA	58	182485	Actinin, alpha 4
GGCTGGGGGC	54	75721	Profilin 1
GCCCAAGGAC	53	195464	Filamin A, alpha (actin-bp-280)
GAAGCAGGAC	51	180370	Cofilin 1 (non-muscle)
TCACCGGTCA	38	290070	Gelsolin (amyloidosis, Finnish type)

TABLE 6.2. TOP 10 GENES ASSOCIATED WITH THE CYTOSKELETON.

As expected the most abundant genes were actins, myosins and tropomyosins. Muscle isoforms of the myosins and tropomyosins were prominent and transgelin is also specific for cells of mesenchymal origin.

6.2.2.2 PROMINENT ECM GENES

Genes associated with the ECM were of particular interest in this study as the disruption of ECM turnover, resulting in net expansion of the ECM, is the histological hallmark of DN. As expected there were components of the ECM specific for basement membrane together with genes associated with the *in vitro* nature of the culture system (TABLE 6.3).

The basement membrane and mesangial matrix specific collagen (type IV) was highly represented in this library. The expression of the fibrillar collagens, types I and III (0.45% and 0.25% respectively), which are not normally found in mesangial matrix, is in agreement with previous studies of mesangial cells grown in culture, and suggests *in vitro* mesangial cells are activated (reviewed in (Davies, '94, Mene et al., '89). Galectin 1 is a beta galactoside binding protein, implicated in the cell cycle and has been described to induce inflammation, apoptosis and proliferation (Rabinovich et al., '99, Moiseeva et al., '00, Rabinovich et al., '00). Currently this gene has not specifically

been described in glomerular mesangial cells but is highly represented in SAGE libraries of epithelial and fibroblast origin. Matrix Gla protein has been described in chondrocytes and smooth muscle cells and is associated with the active inhibition of ECM calcification in transgenic mouse models (Luo et al., '97). Present at a relatively high abundance in this SAGE library (0.18%), expression of Gla in NHMC supports the role of the mesangia as a modulator of elastic glomerular ECM production and remodelling. Gla is found in comparable levels in SAGE epithelium and prostate libraries. Chondroitin sulphate is a proteoglycan that has been described in many cell systems including mesangial cells and is a constituent of the ECM and cell surface (Alberts et al., '94, Elbein and Kaushal, '99).

ECM			
Tag Sequence	Total	UniGene ID (Hs.)	Gene Description
ATGTGAAGAG	424	111779	SPARC (osteonectin)
ATCTTGTTAC	255	287820	Fibronectin 1
GACCGCAGGA	248	119129	Collagen, type IV, alpha 1
GCCCCCAATA	181	227751	Ggalectin 1
GTTTATGGAT	88	279009	Matrix Gla protein
GATGAGGAGA	67	179573	Collagen, type I, alpha 2
TGCAATATGC	49	750	Fibrillin 1 (Marfan syndrome)
TTAGTGTCGT	47	111779	SPARC (osteonectin)
GAGCCTGGAT	42	9004	CSPG 4
GATCAGGCCA	37	119571	Collagen, type III, alpha 1

TABLE 6.3. GENES ASSOCIATE WITH THE ECM.

Fibronectin-1 and collagens were highly represented and SPARC was the most abundant gene in the subgroup. Most have been previously identified in mesangial cells and the presence of type I and III collagens is indicative of *in vitro* culturing.

6.2.2.3 PROMINENT TRANSCRIPTION AND TRANSLATION FACTORS

Many of the highest abundance genes were ribosomal proteins and transcription factors, but many were removed from the transcriptome due to tag redundancy in the mapping. Even so, the ribosomal genes remain in particular high abundance (TABLE 6.4).

Ribosomal protein S27 (RPS27) was the highest abundance in the 2° transcriptome and is up regulated in TGFβ stimulated mammary carcinoma cells in the presence of cycloheximide (Fernandez-Pol et al., '93, Santa Cruz et al., '97)). The gene is also significantly up regulated in malignant and transformed cells when compared to

normal tissue or primary cultures. Originally termed metallopanstimulin-1 (MPS-1), the cDNA for MPS-1 and RPS27 differ by 1bp in the polyadenylation signal (Tsui et al., '96).

Transcription & Translation Factors			
Tag Sequence	Total	UniGene ID (Hs.)	Gene Description
Translation			
CACAAACGGT	266	195453	Ribosomal protein S27
AAGACAGTGG	199	296290	Ribosomal protein L37a
AGGCTACGGA	149	119122	Ribosomal protein L13a
AGCACCTCCA	143	75309	ETL factor 2
ATTCTCCAGT	136	234518	Ribosomal protein L23
GCATAATAGG	125	184108	Ribosomal protein L21
CGCCGCCGGC	123	182825	Ribosomal protein L35
CTGGGTTAAT	114	298262	Ribosomal protein S19
TCAGATCTTT	112	108124	Ribosomal protein S4, X-linked
ACATCATCGA	105	182979	Ribosomal protein L12
Transcription			
ATAGAGGCAA	24	173714	MORF-related gene X
TTCCGGTTCC	18	172609	Nucleobindin 1
ATCCGGCGCC	16	172772	TEF B (SIII)
GGCCCTAGGC	16	78909	Zinc finger protein 36
ATAGACGCAA	15	6353	MORF-related gene 15
CGCACCATTG	15	94672	GCN5-like 1
AGCCCTCCCT	13	74111	RNA-binding protein
GTGGCATCAC	13	14317	Nucleolar protein A3
TGTAATCAAT	13	249495	hNRP A1
TAGATTTCAA	12	197540	Inducible factor 1, alpha subunit

TABLE 6.4. GENES ASSOCIATED WITH TRANSLATION (RIBOSOMAL PROTEINS) AND TRANSCRIPTION.

Many of these genes are present at high abundance but tag redundancy resulted in their removal from the table. This may indicate that the genes are highly transcribed and are either quite inaccurate in transcription or represent multiple transcript variants.

6.2.2.4 PROMINENT METABOLIC ENZYMES

The majority of enzymes, whether metabolic or otherwise are represented in lower abundance classes (TABLE 6.5). As expected, GAPDH features in the metabolic enzyme group together with other genes of involved in or in side reactions in glycolysis. Ornithine decarboxylase antizyme 1 inhibits the biosynthesis and uptake of polyamines by binding to ornithine decarboxylase (Hayashi et al., '97, Nilsson et al., '97). Polyamines are implicated in mitotic spindle formation and chromatin condensation and ornithine decarboxylase levels are increased in the G_2/M and G_1/S transition points in the cell cycle (Pyronnet et al., '00). Such high levels of antizyme 1 probably indicate the arrested nature of the cells in the culture protocol, i.e. the NHMCs were serum starved and showed markedly reduced proliferation. Calpain 1 is an intracellular, Ca^{++}

dependent protease. Various members of the oxidative phosphorylation pathways are well represented in this enzyme group with cytochrome oxidase and reductase subunits represented in the top 10 enzymes. Transcription enzymes, such as RNA polymerase II subunits, were also well represented in this group.

Enzymes			
Tag Sequence	Total	UniGene ID (Hs.)	Gene Description
TACCATCAAT	157	169476	GAPDH
TTGTTGTTGA	73	182278	Calmodulin 2
TTGTAATCGT	51	125078	Ornithine decarboxylase antizyme 1
TTGGAGATCT	48	50098	NADH DH 1 alpha4
CCCCAGTTGC	43	74451	Calpain, small subunit 1
ACTGGGTCTA	41	275163	non-metastatic cells 2, protein
GGCCCTGAGC	38	71618	RNA Polymerase II L
AATATGTGGG	33	351875	Cytochrome c oxidase subunit VIc
TGAGGGAATA	33	83848	Triosephosphate isomerase 1
GGGCGCTGTG	32	8372	Ubiquinol-cytochrome c reductase (6.4kD) subunit

TABLE 6.5. TOP 10 ENZYMES IN THE 2° TRANSCRIPTOME.

As predicted GAPDH is highly represented as are genes involved in oxidative phosphorylation and transcription. High levels of ornithine decarboxylase antizyme 1 indicate the arrested cell cycle.

6.2.2.5 PROMINENT RECEPTORS AND ANTIGENIC MARKERS

Various integrins are represented in high abundance, as are HLA antigens (TABLE 6.6). Integrins are receptors for the ECM and are capable of initiating intracellular signal cascades based on the ECM to which they bind (Elbein and Kaushal, '99, Brady et al., '00, Alberts et al., '94). CDC 151 forms part of the hemidesmosomes, which are the structures through which cells attach to basement membranes. Various receptors for cytokines and other growth factors were present in this NHMC library, but their abundance precluded them from the top 10. Of particular note are the oxytocin receptor (0.1%) and endocytic receptor (0.09%), the receptors for oxidised LDL (0.03%), ANP, parathyroid hormone, transferrin, TNFα and retinoic acid.

Receptors/Antigens			
Tag Sequence	Total	UniGene ID	Gene Description
TGCCTCTGCG	151	75564	CD151 antigen
GTTGTGGTTA	115	75415	Beta-2-microglobulin
CTGACCTGTG	32	77961	MHC, class I, B
TAACTTGTGA	26	295726	Integrin, alpha V (CD51)
TGAAGTTATA	26	287797	Integrin, beta 1 (CD29)
CGACGAGGAG	26	9999	Epithelial membrane protein 3
AACTAATACT	19	295362	DR1-associated protein 1
GTACTGTAGC	18	265829	Integrin, alpha 3 (CD49C)
TACTTGTGTG	17	6354	Stromal cell derived factor rc1
ATCACACAGC	17	79386	Leiomodin 1 (smooth muscle)

TABLE 6.6. TOP 10 RECEPTORS OR ANTIGENIC MARKERS IN THE NHMC TRANSCRIPTOME.

Highly represented are the integrins, beta1 (fibronectin receptor) and alpha V (vitronectin receptor). Cell adhesion and connecting protein were also represented and smooth muscle specific leiomodin 1.

6.2.2.6 PROMINENT CYTOKINES AND CELLULAR FACTORS

Many factors previously described in MC were present in this library but their low abundance precluded them from this list. Of the most abundant genes, many have been described in MC and elsewhere (TABLE 6.7).

IGFBP-7 is low affinity member of the IGFBPs, having 5-6 fold lower affinities to IGF-1 than IGFBP3 (Oh et al., '96). CTGF was also considered to part of the same IGFBP family as IGFBP 7, although this is contentious (Grotendorst et al., '00). Both these genes are present in high abundance in NHMC. In normal kidney fibroblasts, CTGF is known to be involved in the modulation of ECM and connective tissue through the action of TGFβ1 (Duncan et al., '99), the cell cycle (Kothapalli and Grotendorst, '00). In addition, CTGF is a member of the CNN family of small highly homologous proteins. Up regulation of CTGF has been described in cultured human mesangial cells cultured in high glucose (Murphy et al., '99). Follistatin-like 1 is a protein that inhibits the release of Inhibitin betaA (Follicle stimulating hormone release protein, Activin A). Both these genes are present in high abundance in this library.

Cytokines & Cellular Factors			
Tag Sequence	Total	UniGene ID (Hs.)	Gene Description
Cytokines			
TTTGCACCTT	260	75511	CTGF
CATATCATTA	232	119206	IGF binding protein 7
CCGTGACTCT	131	296267	Follistatin-like 1
ACTGAGGAAA	58	77326	IGF binding protein 3
ATCGTGGAGG	43	727	Inhibitin, beta A (activin A)
GGTTGGCAGG	38	3745	Milk fat globule-EGF factor 8 protein
CTTTGAACGA	37	75511	CTGF
AACGCGGCCA	35	73798	Macrophage migration inhibitory factor
AAATGCCACA	32	65450	Reticulon 4
GATAACTACA	24	119206	IGF binding protein 7
Cell Cycle			
GGCTGTACCC	71	108080	Cysteine and glycine-rich protein 1
CTGAGAGCTG	31	78501	Growth arrest-specific 6
TGGTACACGT	25	279574	Cell death-regulatory protein GRIM19
GCACAAGAAG	24	289721	Growth arrest-specific 5
AAAGTCTAGA	21	82932	Cyclin D1
CTTGATTCCC	17	77266	Quiescin Q6
CCACTCCTCA	12	82890	Defender against cell death 1
GCCTCTTGAA	10	184326	CDC10
ACTATTAGTG	10	3436	CDK2-associated protein 1
ACTAACTGTG	10	16003	Retinoblastoma-binding protein 4

TABLE 6.7. TOP 10 GENES FOR CYTOKINES AND THOSE ASSOCIATED WITH CELL CYCLE.

Many of these genes have previously been described in MC. Of particular note is CTGF, which is present three times at high abundance (n = 260, 37 and 14).

6.2.2.7 MISCELLANEOUS GENES

Other genes of interest present at high abundance were protease inhibitors, particularly PAI-1, together with other members of the SERPIN family. Heat shock proteins, in particular the HSP70 and members of signal transduction pathways (TABLE 6.8) are present. Scaffold proteins, classically recognised as containing LIM domains, were also present at high abundance. These proteins generally interact with the cytoskeleton and help to juxtapose various enzymes, co-factors and other proteins to an active site (Brown et al., '99).

Misc			
Tag Sequence	Total	UniGene ID (Hs.)	Gene Description
CCCAAGCTAG	100	76067	Heat shock 27kD protein 1
TTGGGGTTTC	82	62954	Ferritin, heavy polypeptide 1
TGCATCTGGT	79	75410	Heat shock 70kD protein 5
ACCTGTATCC	51	182241	Interferon induced transmembrane protein 3 (1-8U)
TAAAAATGTT	44	82085	SERPIN E (nexin, PAI type 1), member 1
CCCCCTGGAT	42	275243	S100 calcium-binding protein A6 (calcyclin)
TAATGACAAT	40	239069	Four and a half LIM domains 1
AGCCTTTGTT	39	9930	SERPIN H (HSP 47), member 2
ACAAGTACCC	36	142827	P311 protein
GGTGGCACTC	36	77273	Ras homolog gene family, member A

TABLE 6.8 TOP 10 GENES NOT PLACED IN ANY OF THE OTHER GROUPS.

This group consists mainly of the Heat Shock Proteins (HSP) and proteases (SERPINS).

6.2.3 GENES OF POTENTIAL FUNCTIONAL SIGNIFICANCE IN DN

Many of the genes thought to have functional significance in diabetic nephropathy (previously implicated in glucose stress or DN) were represented by tags below the threshold frequency for the top10/20 tables, but are collated in TABLE 6.9. While most of the genes identified as contributing to DN are represented in this NHMC library, most were of low frequency or masked by tag and gene redundancy and so could not be accurately quantified. They remain, however, of high enough abundance to be accurately detected.

Many of the high abundance genes present in these tables are members of gene families whose members or isoforms have been characterised in a cell or tissue specific manner. Many of the cytoskeletal and ECM genes such as the myosins, collagens and actin interacting proteins are isoforms characterised in smooth muscle. The gene for actin could not be resolved in this SAGE analysis, but the PCR product generated for the abundance experiments was sequenced and shown to be the smooth muscle isoform of actin (α2-SMA). Mesangial cells and smooth muscle cells are believed to share common myoblast heritage and the presence of these smooth muscle specific isoforms in MC supports this.

Signal transduction pathways were also represented in this library. Many mitogen activated protein kinases (MAPKs) were mapped and five isoforms of PKC were identified. All detected PKC isoforms were represented at low levels (1-2 tags). PKCβ1, widely cited as the prominent form of PKC in MC, was present at the same levels as PKC ν (nu), ι (iota), ζ (zeta) and μ (mu). The 1° transcriptome revealed that the tag mapping to PKC β1 is ambiguous for one other gene, while the other PKC isoforms have reliable mapping. Interestingly, two tags mapped reliably to PKC ζ, suggesting a transcript ambiguity.

Tag	Freq	UniGene ID	Gene	Function
*AAGATCAAGA	445	Hs.195851	α2-SMA	Cytoskeleton
GGAGTGTGCT	232	Hs.9615	Myosin lc2	Cytoskeleton
TTGCACCTT	131	Hs.75511	CTGF	Cytokine
*TGGAAAGCTT	2	Hs.77202	PKC β1	Signal Transduction
AAATTAATTG	2	Hs.143460	PKC ν	Signal Transduction
TCAAAATTTA	2	Hs.1904	PKC ι	Signal Transduction
CGCATTAAAG	2	Hs.78793	PKC ζ	Signal Transduction
TTATATTTAA	1	Hs.2891	PKC μ	Signal Transduction
TTGGGTATCC	2	Hs.1674	GFAT	Hexosamine Pathway
*AAGAGTTTGA	1	Hs.75313	AR	Polyol Pathway
GGGGCTGTAT	3	Hs.1103	TGF β1	Cytokine
TTATGTATCA	6	Hs.169300	TGF β2	Cytokine
TGCCACACAG	1	Hs.2025	TGF β3	Cytokine
TTTTGTGCAT	3	Hs.238990	$p27^{kip1}$	CDK Inhibitor

TABLE 6.9. GENES OF FUNCTIONAL SIGNIFICANCE IN DN.

Previously characterised genes described in MC that are present in the SAGE library. '*' Indicates that the tag mapping to this gene cluster is ambiguous for another gene.

The pro-fibrotic cytokine TGFβ1 has been intensely studied in MC. A prominent role for TGFβ1 in the development and progression of DN has been emerging over the last decade. Interestingly the β2 and β3 isoforms are also present in this library. Recent reports have indicated the separate and sequential role for TGF β isoforms in the development of fibrotic scaffolds preceding calcification in differentiating osteoclasts (Nugent et al., '98, Gosain et al., '00).

The CDK inhibitor $p27^{kip1}$ has also been described in MC and in pathology DN (Wolf et al., '97). This protein specifically inhibits cyclins and halts the progression of the cell cycle at the G1/S checkpoint (Polyak et al., '94). The activity of $p27^{kip1}$ was shown to be TGFβ dependent as treatment of cells arrested in G1/S by high glucose culture was reversed with the addition of anti-TGFβ antibody (Wolf and Ziyadeh, '99).

The rate-limiting enzyme in the hexosamine pathway, GFAT, was also present in this library albeit at a low level (2 tags). The rate limiting enzyme in the polyol pathway, AR, is also present in this SAGE library although the tag for AR is ambiguous.

6.2.4 SUMMARY OF THE NHMC TRANSCRIPTOME

The 2° transcriptome was classified into functional groups. As expected genes of the cytoskeleton, translation and transcription factors were the most abundant genes. Several structural genes showed isoform specificity to smooth muscle cells and some genes had been characterised exclusively in SMC (e.g. TAGLN). Genes that have previously been characterised in primary mesangial cells were also present in this SAGE library such as types I, III and IV collagen. Type I and III collagens indicate activation of MC due to *in vitro* culturing. Another gene that indicates activation of MC is a high level of α2 SMA.

Genes associated with DN were also present in this NHMC library, such as CTGF, Collagen isoforms, FN-1, $p27^{kip1}$, TGFβ1, PKC isoforms and metabolic enzymes GFAT and AR. Many of these genes were present at low levels and so precludes them from accurate analysis, particularly for determining differential transcription, as will be discussed in CHAPTER 7.

6.3 TAG ANOMALIES

6.3.1 AMBIGUITY IN TAGS

Mapping tags to genes revealed errors in tag assignment and gene representation. This redundancy in mapping can be a consequence of two phenomena. First is the presence of a conserved region in several genes, examples of which are a shared exon or signal region. A second possibility is the technical failure of SAGE. The most likely cause for the generation of the same tag mapping to many genes was a high degree of adenine in the tag. This would result from the most 3'N*la* III site being adjacent to a region rich in adenine residues. This region may be a polyA tail or any

other region than can serve as a priming site for the initial cDNA synthesis. There appears little use in attempting to discriminate between the genes whose tags are rich in adenine or tags rich in other bases that indicate repetition elements such as *Alu*. The second more interesting reason for tag ambiguity could be tag generation from a region conserved or indicating specific function. This is an attractive theory as it could form a way of linking previously unrelated genes in a structural manner. For example, if 30 separate genes generate the same tag, is this tag generated from conserved sections of the genes and if so could this conserved region indicate a shared functional site? Resolving this was considered to be outside the scope of this project.

6.3.2 AMBIGUITY IN GENES

It does not always follow that a single tag represents a single gene. This observation may be a powerful tool in determining the transcription dynamics of a gene. From the estimation of errors for contamination, it would appear that this SAGE library is of generally high quality and so favours the identification of alternative transcripts. Supporting evidence for this phenomenon comes for the high-density clusters of ESTs generated from the HGMP. By investigating the specific locus for the gene, it may become clear that there are different transcripts from the same gene. To test this theory two genes are used. One gene selected from the literature, has previously described alternative transcription (RTN4) and one gene, from the SAGE library, predicts some sort of alternative transcription not previously described (CTGF).

6.3.2.1 RTN4

The reticulon gene family is a group of neuroendocrine specific genes, characterised mainly in neurons but present in NHMC. Reticulon 4 (RTN4) is protein whose primary function appears to involve growth inhibition in neural cells (Spillmann et al., '98). Axon regeneration is generally successful in the peripheral but not in the central nervous system. The inhibition of neural outgrowth is caused by several factors, one of which is RTN4 (previously nogo or inhibitor of neural outgrowth) (Chen et al., '00). Cloned RTN4 suggests that there are 3 splice variants A, B, and C, and these are generated from post-transcriptional splicing. Variant A is 1,192 amino acid residues, variant B and C lack residues 186-1,004 but variant C also has a smaller amino-terminal. This data implies that all the variants are derived from the same transcript and

will have the same carboxy terminal. SAGE analysis should be of little use in distinguishing between the variants, yet the SAGE tag mapping revealed three major tags. One tag maps reliably to two mRNA sequences and two further tags map to two clusters of ESTs. All three tags fail to map to any other genes.

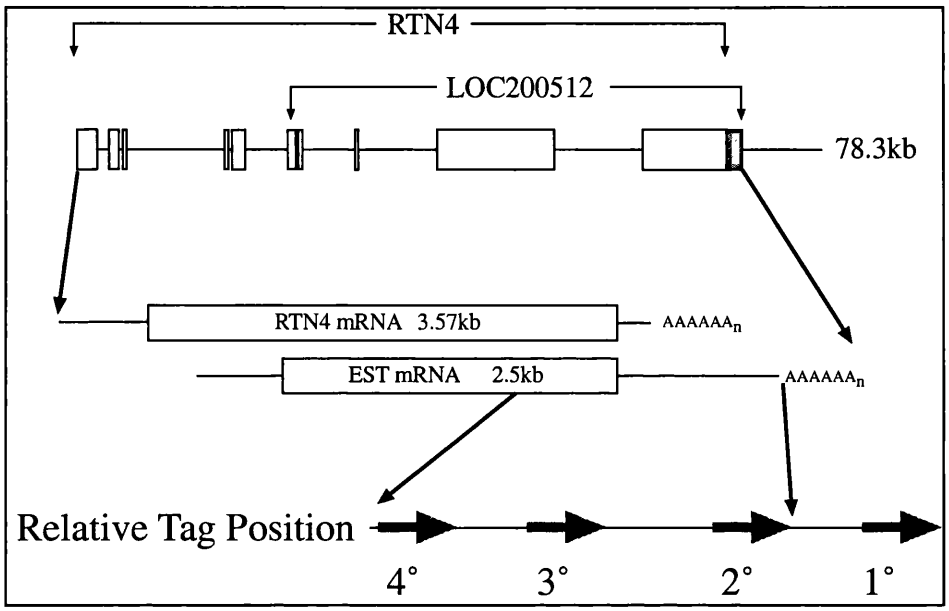


FIGURE 6.1. TAGS GENERATED FROM THE RTN4 GENE.

The RTN4 locus contains two genes; RTN4 and LOC200512 11 exons (boxes) are distributed across 78kb. LOC200512 begins transcription approximately half way into RTN4 transcription but extends 582bp beyond RTN4 transcription termination. The 1° Tag maps to mRNA reference sequences but fails to map to EST clusters. The 2° and 3° tags map to about equal numbers of ESTs in two clusters; one is about 600bp longer and consensus sequences match near 100% in nucleotides and code in the same frame. The 4° tag is also well represented in SAGE library, and may indicate a further transcript from this locus. (Image not drawn to scale and the shaded area on the DNA represents the extra 600bp of 3' transcription)

Inspecting the locus containing RTN4 (LocID 57142) reveals that there are two annotated genes at this site, RTN4 and LOC200512. Transcription of LOC200512 mRNA begins 1.8kb after the initiation of transcription of the RTN4 mRNA, and extends for some 600bp past the RTN4 termination of transcription; this suggests two transcripts with 100% homology over 1200bp. Where they overlap their coding sequences are in frame, suggesting shared exons (see FIGURE 6.1). Two mRNAs are generated from this locus and SAGE has predicted this. Further interrogation of the transcript variants is required to determine any functional significance of the alternative transcripts. A final tag maps to 13% of the cluster sequences. This tag appears to be generated from the 7th of 11 exons, and so may represent an additional transcript (or splice variant) from this locus.

The primary RTN4 tag has mapped reliably to two characterised mRNA sequences but this tag has not been supported by any EST cluster data, which according to 'SAGEmap', convincingly map to two tags at equal levels (see TABLE 6.10). The primary tag is not seen in the NHMC library and the failure of the mRNA to map to the tags appears to be due to either an error in the reference sequence data or a polymorphism in the anchoring enzyme site or tag. As there are currently no indications of SNPs, the primary tag may have been generated from a mistake or tag extraction error. The EST clusters suggest that there are two variants. Both EST sets share high homology but one set appears to have an extended 3' UTR. From this evidence it would appear that one tag represents the actual RTN4 transcript and the other high abundance tag represents the EST highly similar to RTN4.

Tag	Tag Sequence	Sequence Clusters	NHMC freq.	Ave SAGE freq (per 10,000)
1°	TGACTGTAAA	2/1385 (0.14%)	-	-
2°	TGTTCATCAT	342/1385 (25%)	18 (36%)	4
3°	AAATGCCACA	308/1385 (22%)	32 (64%)	3.3
4°	TGAACTGCAC	180/1385 (13%)	-	2

TABLE 6.10. TAGS IN NHMC LIBRARY THAT MAP TO RTN4 (HS.65450).

The primary tag was not present in the NHMC library or any other library and is possibly due to incorrect tag extraction. The 2° and 3° tags are present at similar levels while the 4° tag was not detected. Average SAGE data was calculated from the libraries used for the virtual northern (CH4.5).

6.3.2.2 CTGF

Connective tissue growth factor is a member of CCN family and was previously known as IGFbp8. A brief description of CTGF was presented previously in this chapter and will not be discussed here.

Tags mapping to CTGF are present 3 times at a frequency greater than 10 (0.56%, 0.08% and 0.05% respectively), suggesting three differentially transcribed or processed mRNAs. All three tags map only to CTGF and are present in other SAGE libraries at similar or greater levels. Unlike the RTN4 locus, the CTGF locus contains only one annotated gene. This information suggests that CTGF tags are a result of alternative transcription or splice variants. Unlike the RTN4 SAGE data, each of these three tags does not show such clear mapping to EST clusters (see TABLE 6.11)

Tag no	Tag Sequence	Seq. Clusters	NHMC tag freq	Average SAGE freq
1°	TTTGCACCTT	378/720 (52%)	260 (84%)	15.3
2° (AS)	CTTTGAACGA	2/720 (0.03%)	37 (12%)	5
3°	AAATGGATAC	42/720 (6%)	14 (4%)	3

TABLE 6.11. TAG CLUSTERS FOR CTGF (HS.75511).

Sequence clusters illustrate the ESTs and mRNA that group to a particular tag. Seq clusters indicate the EST clustering information from 'SAGEmap'. The average SAGE freq is a mean frequency across the ten SAGE libraries used for a virtual northern where the tag is present. Note that the 2° tag is oriented in the antisense (AS) relative to the mRNA.

Of the three tags, two mapped to a characterised mRNA and examination of the representative cDNA sequence suggests a premature transcription termination prior to the most 3'N*la* III site (see FIGURE 6.2). This termination resulted in a shift of the most 3' N*la* III site from –64bp (relative to the poly-A tail) to –333bp. Mapping the third tag proved more complicated. The origin of this tag was again assigned to the 2nd most 3' N*la* III site at –333bp relative to the poly-A tail, but in the opposite direction. The mapping information for this tag suggests two possibilities. Firstly, that the tag resulted from technical failure in the SAGE procedure, namely the presence of contaminating 5' digestion products prior to linker addition. This seems unlikely, as this phenomenon does not occur with other high abundance genes. Secondly, the tag may be real, generated from differentially transcribed or processed mRNA but transcribed or processed in such a way as to facilitate SAGE tag generation in the opposite direction to the previous two tags.

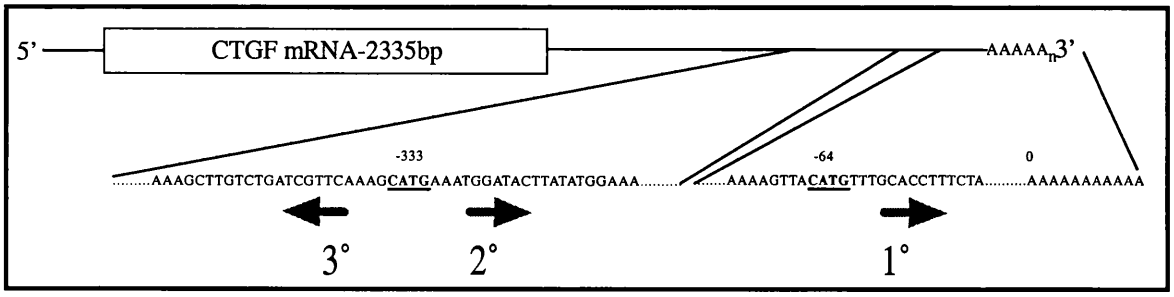


FIGURE 6.2. TAGS GENERATED FROM THE CTGF GENE.

Only one gene is annotated at this locus. The primary tag satisfies the SAGE criteria, the 2° tag appears to result in a shift in position of the anchoring enzyme and the 3° tag appears to be generated from an antisense strand. Arrows indicate tag position and orientation.

There is compelling evidence that the 3'UTR of the CTGF mRNA contains regulatory elements that are active in both sense and anti-sense configurations (Kubota

et al., '99, Kubota et al., '00). Clearly this provides evidence for a more through interrogation of CTGF gene transcription and tag assignments.

6.3.3 SUMMARY OF TAG ANOMALIES

From the above *in silico* experiments, it appears the SAGE analysis can offer insights into the transcription of genes. The literature contains widespread reports of alternative transcription or splicing and SAGE analysis may offer a way of identifying and classifying any such transcription variation. The locus for RTN4 annotates two genes that also cluster two sets of ESTs, suggestive of alternative transcription or overlapping genes. The NHMC SAGE library contained high levels of RTN4 tags, was sensitive to this heterogeneity, also identifying two tags that mapped to the same UniGene cluster but different EST clusters. The locus for CTGF annotated only one gene, but SAGE analysis predicted three tags that reliably map to this gene, suggesting alternative transcripts from a single gene. Evidence in the literature also suggests alternative transcription and *cis*-acting elements that affect transcription. These experiments have identified two ways where SAGE can be useful in identifying alternative transcription.

6.4 VIRTUAL NORTHERN

6.4.1 COMPARING TRANSCRIPTOMES

The digital nature of SAGE facilitates comparison between libraries generated in independent experiments. In comparing libraries sourced from other tissues it is possible to investigate and identify genes that are present in all libraries and those that are transcribed in a restricted pattern. High abundance housekeeping genes, which are predicted to be present at high levels across the entire database, should confirm this. In addition to housekeeping genes, specific genes indicating a particular cell type should only be present in libraries derived from that lineage. Together with characterised genes, there could also be un-described genes that share a restricted pattern of transcription.

The NHMC library was compared to other SAGE libraries to assess it's validity by predicting and testing the presence of common or housekeeping genes. Together with this validation, analysis of restricted patterns should reveal tags that have a particular association to NHMCs or cells of a similar linage. This hypothesis was generated from the comparison experiments in CHAPTER 4.3, where a clear and positive correlation was observed between tags generated from cells of similar linage. The theory implies that cells from a similar linage or of similar phenotype will express similar genes. In order to investigate this further a digital Northern was generated, where tag frequencies were normalised across a selection of SAGE projects.

6.4.2 CONSTRUCTING A VIRTUAL NORTHERN

A virtual Northern database was compiled from SAGE data available at the time of the experiment. SAGE files were downloaded form SAGEmap and normalised tag frequencies uploaded onto the prepared database. The database was designed to show all tags present, and thus the ability to interrogate tag frequencies across different libraries, from different cell types (see TABLE 6.12).

Library	Cell type	Source	Library ID		Total tags
			dbEST Lib ID	CGAP Lib ID	
293	Cell line	Embryonic Kidney	4912	538	43527
Cerebellum	Normal	Brain	4036	-	51280
Kidney	Normal (Bulk)	Kidney	9707	673	41857
Fibroblast	Cell line	Skin	-	650	8851
H126	Epithelium	Pancreas	5636	560	31512
NHA	Astrocytes	Brain	1803	355	52261
Heart	Normal (Bulk)	Heart	10076	-	84357
Liver	Normal (Bulk)	Liver	10051	690	66861
Prostate	Normal (Bulk)	Prostate	7294	642	13302
PR317	Normal (Dissected)	Prostate	5639	570	59513
NHMC	Normal (Primary cell)	Kidney	-	-	43358

TABLE 6.12. LIBRARIES USED IN THE CONSTRUCTION OF VIRTUAL NORTHERN DATABASE.

Their source is indicated as normal (primary) or a cell line and dbEST (ncbi) and CGAP (cancer genome anatomy project) library ID's listed. The total tags sampled in each library were included to estimate the complexity of the library. All libraries were normalised to 10,000 tags.

Three queries generated three 'digital Northerns'. First, a search was made comparing the frequency of all tags across all libraries, which constitutes the complete digital Northern. The second search was for tags present only in the NHMC library, suggesting highly restricted expression of genes. Finally, a search was made for genes that were expressed predominantly in NHMC. A schema of the database is shown in FIGURE 6.3.

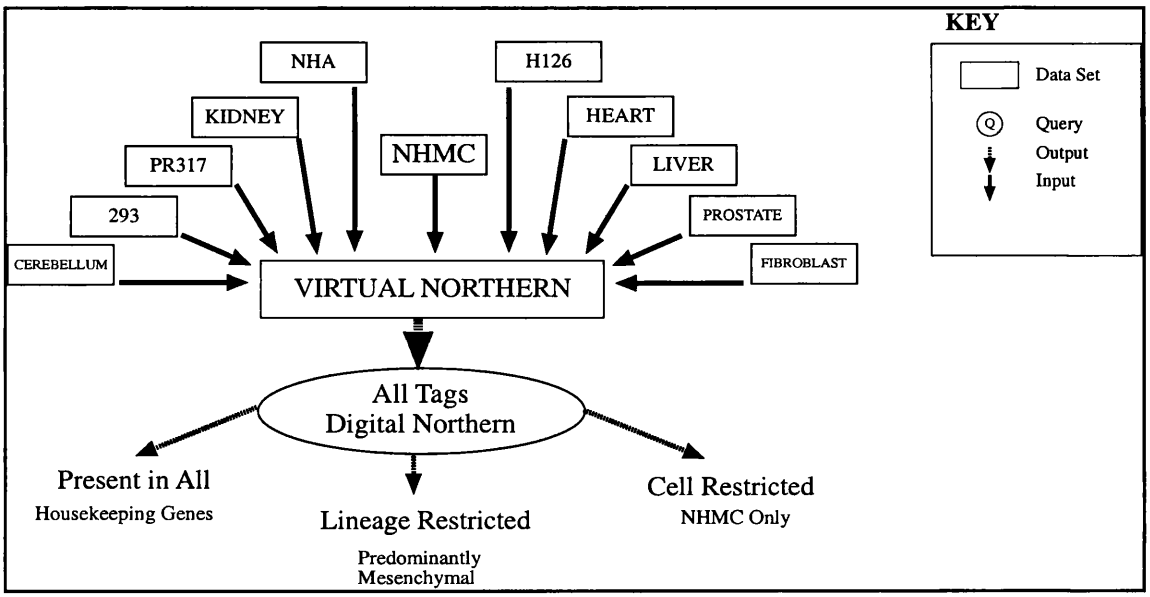


FIGURE 6.3. SCHEMA OUTLINING CONSTRUCTION AND OUTPUT FROM THE VIRTUAL NORTHERN.

Information of the libraries used can be found in TABLE 6.12. Where possible libraries obtained from normal tissue or primary cells was preferred so to avoid transformation elements.

6.4.2.1 HOUSEKEEPING GENES (PRESENT IN ALL)

a

HOUSE KEEPING GENES

Tag Sequence	293	Cere	Fibro	H216	Heart	Kidney	Liver	NHA	NHMC	PR317	Prostate	UniGene ID	Gene
AAGATCAAGA					26			14	110	31		1288	ASMA
CACAAACGGT	39	10	31	63	13	7	7	9	59	37	8	195453	RPs27
TACCATCAAT	34	13	9	4	43	4	17	54	35	27	2	169476	GAPDH
AGCACCTCCA	23	13	20	36	15	8	11	24	32	30	7	75309	EF2
TGCATCTGGT			10	7	5		13	4	18	3	4	75410	HSP70
GTGAAACCCC	30	112	28	55	8	43	8	29	28	5	47	44396	Coronin
TTGGGGTTTC	4	4	34	4	38	47	17	43	18	31	17	62954	Ferritin H
GTTGTGGTTA	2		25	6	23	9	30	4	26	45		75415	b2 Microglobulin
GGCAAGCCCC	10	3	23	27	11	8	7	5	13	23	32	76067	HSP27
AGCCCTACAA	4	54	5	21	107	88	21	16	15	60	36	95243	TEF-A
CTGGGTTAAT	46	4	52	13	15	3	14	14	25	47	77	101047	TEF-3
AATCCTGTGG	27		8	25	6	5	3	6	12	15		111334	Ferritin L
ACTTACCTGC	5	6		3	18	4	6	3	6	5	2	174031	CyC Ox Ivb
GAAGCAGGAC	6	7	43	7	5	3	3	25	11	7	7	180370	Cofilin
TGTGTTGAGA	82	15	30	72	7	22	9	40	38	68	14	181165	EF1a
CGCTGGTTCC	15	9	22	26	8	7	10	12	14	14	9	181165	EF1a
GTTCGTGCCA	9	4	15	23	9	3	3	4	17	8	10	195464	Fillamin
TGGGCAAAGC	27	7	15	27	10	10	6	14	9	23	22	2186	TEF1gamma
GTGCTGAATG	2	3	27	5	6	13	5	15	73	43	1	77385	Myosin lc6

b

RIBOSOMAL PROTEINS

Tag Sequence	293	Cere	Fibro	H216	Heart	Kidney	Liver	NHA	NHMC	PR317	Prostate	UniGene ID	Gene
GAGGGAGTTT	18	10	42	22	27	10	15	8	19	29	265	76064	L27a
GTGAAGGCAG	15	10	19	27	12	4	8	3	9	10	15	77039	S3A
AGCTCTCCCT	17	4	28	35	7	6	5	8	18	9	19	82202	L17
TCAGATCTTT	55	7	19	18	12	8	6	11	25	35	10	108124	S4
CCAGAACAGA	23		15	38	17	4	10	7	29	25	17	111222	L30
GGCAAGAAGA	17	3	13	11	7	5	3	4	11	12	4	111611	L27
AATAGGTCCA	11	9	5	11	18	4	10	6	10	18	3	113029	S25
AGGCTACGGA	17	10	31	31	19	10	15	19	33	34	87	119122	L13a
ACTCCAAAAA	5	7	13	2	8	3	4	5	6	10	26	133230	S15
TAATAAAGGT	37	7	2	8	22	4	12	2	15	22	5	151604	S8
ATCAAGGGTG	14	4	6	15	9	3	3	2	7	14	2	157850	L9
TTCAATAAAA	18	3	6	2	38	7	20	18	16	47	24	177592	P1
AATCCTGTGG	27		8	25	6	5	3	6	12	15		178551	L8
CGCTGGTTCC	15	9	22	26	8	7	10	12	14	14	9	179943	L11
GCCTGTATGA	23	3	9	12	4	4	4	5	15	12	8	180450	S24
CCCGTCCGGA	33	10	39	64	6	16	6	12	22	36	56	180842	L13
CCAGTGGCCC	6	3	15	13	2	3	2	2	4	10	2	180920	S9
CGCCGCCGGC	13	10	81	74	11	11	8	5	27	12	110	182825	L35
AAGGAGATGG	21	4	9	37	9	6	6	13	42	20	2	184014	L31
CACAAACGGT	39	10	31	63	13	7	7	9	59	37	8	195453	S27
ATTCTCCAGT	32	6	14	12	13	3	11	15	30	16	4	234518	L23
TAAAAAAAAA	5	4	23		5	7	10	3	3	8	12	244621	S14
GAACACATCC	26	3	6	16	6	6	4	6	12	16	2	252723	L19
TGGTGTTGAG	26	6	18	58	9	5	4	8	20	33	17	275865	S18
CTGGGTTAAT	46	4	52	13	15	3	14	14	25	47	77	298262	S19

TABLE 6.13 A & B. DIGITAL NORTHERN OF HOUSEKEEPING GENES AND RIBOSOMAL PROTEINS (RP).

All digital northern are base of the 2° Transcriptome and so multiple gene mapping is accepted. Several tags may represent the same gene but only the major tag in the NHMC library was considered. This results in the digital northern being insensitive to lineage restricted alternative transcription where it exists.

High abundance genes, particularly the ribosomal proteins, cytoskeleton and transcription factors are present across all the libraries tested (see TABLE 6.13 a & b.) There are some variations with the levels of each particular gene but there is general agreement that these are constitutively expressed. These genes are generally regarded as housekeeping genes as they are used across experimental samples and various other

groups where normalisation between samples is required. Several metabolic enzymes are also regarded as housekeeping genes and are also seen at similar levels across the digital Northern, for example GAPDH.

6.4.2.2 RESTRICTED TRANSCRIPTION IN NHMCS

The digital Northern reveals that some tags, and thus presumably genes have patterns of expression restricted to NHMCs as they are not detected across the libraries contained in this database (TABLE 6.14). These included laminin S, activin A and 4 EST clusters. Interestingly, laminin S and activin A (Inhibitin Beta A) are represented by other tags and so they may indicate cell-restricted isoforms of these genes. The EST clusters may represent uncharacterised genes that also show a restricted expression

NHMC RESTRICTED

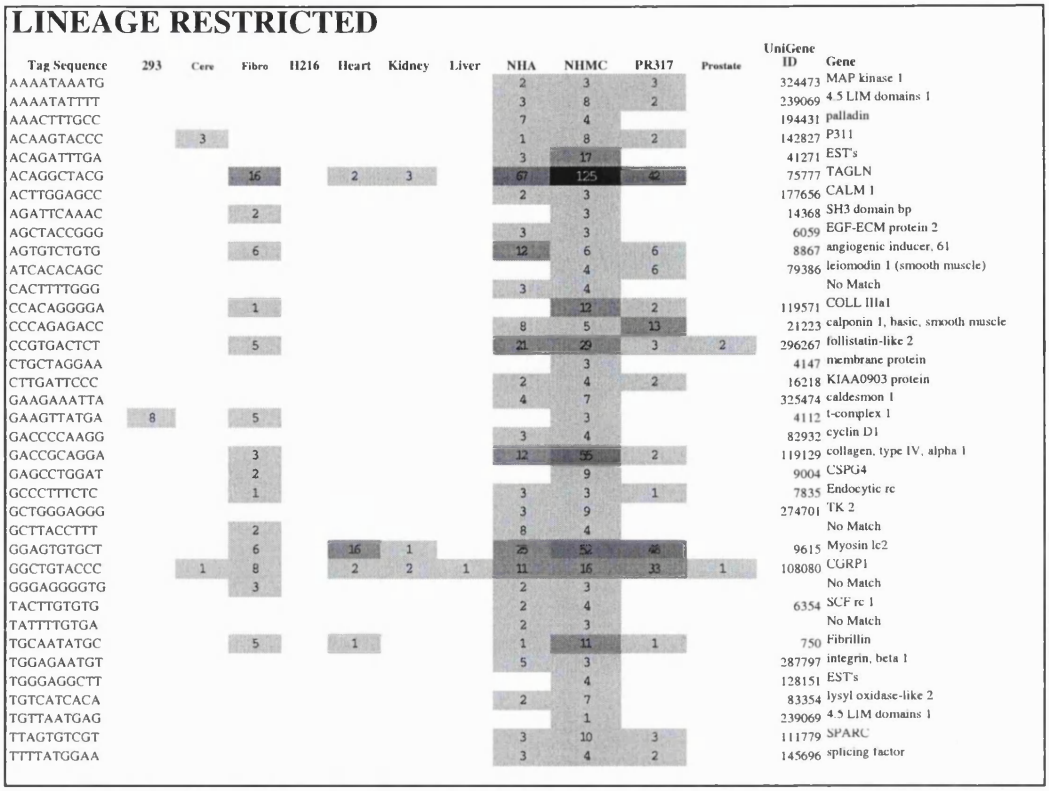
Tag Sequence	NHMC	UniGene ID	Gene
ACTTGGAGTC	13	296842	EST
ATCGTGGAGG	10	727	Inhibitin, Beta A
GTGAATGCCC	10	44165	EST
TCAGACAAAA	5	66881	dynein, cytoplasmic, intermediate polypeptide 2
TCTGCCTATG	5	90291	laminin, beta 2 (laminin S)
TACCTCTCTA	4	296842	EST
TGGGAGGCTT	4	128151	EST
TATGAGGGTA	4	24950	regulator of G-protein signalling 5
GGCCCTAGGC	4	78909	zinc finger protein 36, C3H type-like 2

TABLE 6.14. DIGITAL NORTHERN OF GENES RESTRICTED TO NHMCS.

It is unlikely that there are strict patterns of gene expression for all of these genes but certainly the abundances were markedly higher than the other libraries. The high number of ESTs suggests that NHMC express a restricted pattern of as yet un-characterised genes.

6.4.2.3 GENES PRESENT IN 'NHMC-LIKE' CELLS

Some tags appeared to be restricted to libraries that were derived form cells of mesenchymal origin (see TABLE 6.15). These libraries were astrocytes (NHA), fibroblasts, smooth muscle cells (PR317) and Mesangial Cells (NHMC). Genes represented in this query include mainly specific isoforms of cytoskeletal genes. In particular are the genes for α2-SMA, transgelin and SM specific Myosin lc2. There were also a number of non-structural genes, which appeared restricted predominantly to cells of mesenchymal origin such as angiogenic inducer 61 and follistatin-like2.



LINEAGE RESTRICTED

Tag Sequence	293	Cere	Fibro	H216	Heart	Kidney	Liver	NHA	NHMC	PR317	Prostate	UniGene ID	Gene
AAAATAAATG								2	3	3		324473	MAP kinase 1
AAAATATTTT								3	8	2		239069	4.5 LIM domains 1
AAACTTTGCC								7	4			194431	palladin
ACAAGTACCC		3						1	8	2		142827	P311
ACAGATTTGA								3	17			41271	EST's
ACAGGCTACG			16		2	3		67	125	42		75777	TAGLN
ACTTGGAGCC								2	3			177656	CALM 1
AGATTCAAAC			2						3			14368	SH3 domain bp
AGCTACCGGG								3	3			6059	EGF-ECM protein 2
AGTGTCTGTG			6					12	6	6		8867	angiogenic inducer, 61
ATCACACAGC									4	6		79386	leiomodin 1 (smooth muscle)
CACTTTTGGG								3	4				No Match
CCACAGGGGA			1						12	2		119571	COLL IIIa1
CCCAGAGACC								8	5	13		21223	calponin 1, basic, smooth muscle
CCGTGACTCT			5					21	29	3	2	296267	follistatin-like 2
CTGCTAGGAA									3			4147	membrane protein
CTTGATTCCC								2	4	2		16218	KIAA0903 protein
GAAGAAATTA								4	7			325474	caldesmon 1
GAAGTTATGA	8		5						3			4112	t-complex 1
GACCCCAAGG								3	4			82932	cyclin D1
GACCGCAGGA			3					12	55	2		119129	collagen, type IV, alpha 1
GAGCCTGGAT			2						9			9004	CSPG4
GCCCTTTCTC			1					3	3	1		7835	Endocytic rc
GCTGGGAGGG								3	9			274701	TK 2
GCTTACCTTT			2					8	4				No Match
GGAGTGTGCT			6		16	1		25	52	46		9615	Myosin lc2
GGCTGTACCC		1	8		2	2	1	11	16	33	1	108080	CGRP1
GGGAGGGGTG			3					2	3				No Match
TACTTGTGTG								2	4			6354	SCF rc 1
TATTTTGTGA								2	3				No Match
TGCAATATGC			5		1			1	11	1		750	Fibrillin
TGGAGAATGT								5	3			287797	integrin, beta 1
TGGGAGGCTT									4			128151	EST's
TGTCATCACA								2	7			83354	lysyl oxidase-like 2
TGTTAATGAG									1			239069	4.5 LIM domains 1
TTAGTGTCGT								3	10	3		111779	SPARC
TTTTATGGAA								3	4	2		145696	splicing factor

TABLE 6.15. DIGITAL NORTHERN OF GENES RESTRICTED TO CELLS OF MESENCHYMAL ORIGIN.

Cells from the NHA (Astrocytes), Fibro (Fibroblasts) and PR317 (Prostate) tissue appear to express a similar collection of genes. These genes include not only 'smooth muscle' isoforms of cytoskeletal and contractile machinery but also some growth factors and receptors.

6.4.3 SUMMARY OF THE DIGITAL NORTHERN

Interrogation of the virtual Northern permits the investigation of the similarities and differences between other SAGE libraries. From CHAPTER 3 it would appear that the majority of genes are present in low abundance and that variation between genes present in cell transcriptomes is minimal. Given this observation there should also be very little difference between high abundance tags present in the digital Northern. This prediction is supported by the observation that high abundance housekeeping genes do not change and only a handful of detected genes fluctuate between libraries.

There appeared to be a restricted pattern of transcription for mesenchymal cells in a number of genes, for example myosin isoforms. Additionally, a catalogue of genes was generated where transcription appears restricted to NHMC only. Genes such as the ECM proteins COLL (type IV and III), CSPG4, FN-1 and laminin b1 appear to be expressed much more in NHMCs than in other libraries. Several ESTs appear to be

restricted exclusively to NHMCs, which may provide a basis for the characterisation of novel genes in the setting of NHMCs.

These experiments provide evidence that *in silico* mining of SAGE data can achieve many objectives. The availability of SAGE data means that many tissues can be screened for particular genes without prior knowledge of sequence or gene characterisation. Using a tag as a gene identifier means that data is minimal and large amounts can be manipulated and shared easily. Mining in this way can identify genes that are expressed across all libraries and genes that are restricted to each particular library. As the phenotype of a cell is defined by the genes it expresses and the manner in which they are transcribed, identifying restricted gene expression will identify transcriptional elements associated with specific phenotype. Mining for restricted expression will also identify genes that can form a basis of classification.

6.5 DISCUSSION

The presence of high abundance genes with specific function in ECM construction and remodelling suggests a specific role for these genes in the context of the mesangial cell and the surrounding matrix. Whether this reflects a role for these genes in normal mesangial cell function or a result of *in vitro* culture remains to be determined. Sub-confluent mesangial cells secrete ECM components, which continues to culture confluence when the cells begin to form hillocks and other 3D structures. Cells within 3D structures display an altered phenotype to monolayers and react differently to stimuli, as described by studies growing mesangial cells in collagen gels (Marx et al., '93, Saito et al., '93, Sugiyama et al., '98). Clearly further investigation into the extracellular environment and its role in cellular growth will provide much information relating to the phenotype of a cell and contribute to a greater understanding of the MC.

Masking of tags and genes is common in all SAGE libraries and is an inescapable feature of SAGE analysis. This occurs because of non-uniqueness of both the tags and the 3'mRNA sequence. Investigating non-unique tags was not considered in this study beyond the extension of the tag to the 11^{th} base. The possibility of one gene being represented by many tags, and tags that may represent differentially transcribed or processed gene products, was further investigated. This particular type of

masking may also occur because of incomplete anchoring enzyme digestion, or miss priming of cDNA synthesis. Being able to distinguish between different mRNA species and incomplete digestion may prove to be a powerful indicator of SAGE library integrity as well as an insight into different products of the same gene. In this study, a simple test extracted all the possible tags from a representative mRNA sequence and then examined the frequency of appearance for each tag. 94% of the tags mapping to a selected housekeeping gene cluster were derived from the correct 3' AE site, and thus mapped correctly. However, mapping tags to 2° or 3° AE sites may not always be an error and may reflect the sensitivity of SAGE to the generation of different gene products. Two genes were used to investigate this, CTGF and RTN4. Both genes had tags that were generated from AE sites 5' to the 1° AE site, and the levels of these secondary tags suggested different transcripts as the phenomenon was not observed in other high abundance genes at such levels.

Tags mapping to CTGF are present 3 times at a frequency greater than 10 (0.56%, 0.08% and 0.05%), which suggests three differentially transcribed or processed mRNAs. All three tags map only to CTGF and are present in other SAGE libraries at similar or greater levels; also suggesting they were derived from alternative transcripts. Of the three tags, two mapped to a characterised mRNA and examination of the representative cDNA sequence suggests a premature transcription termination prior to the most 3'N*la* III site. The origin of the third tag was again assigned to the 2nd most 3' N*la* III site at –333bp relative to the polyA tail, but in the opposite direction. Absence of 5' digestion products in other genes suggests the tag is real, generated from antisense transcripts. From the data generated in this SAGE analysis, it would appear that CTGF has several transcriptional elements associated with the gene and requires more through investigation.

RTN4 mapped to two major tags both of which group with two distinct EST clusters. The Locus for RTN4 currently has two characterised mRNAs present, overlapping at the 5'end. One transcript extends a further 600bp beyond the termination of the other. Investigating the position of the two tags mapping to RTN4 reveals that each tag represents a transcriptional product from this gene, one containing an extra 600bp 3' tail, which contains an extra AE site. The fact that these tags appear at about equal frequency suggests that neither gene product is favoured over the other and may indicate different functions of very similar transcripts.

As the total population and complexity of SAGE libraries grow it should become more apparent which gene variants are more common and which are not so common. Clearly this type of analysis is currently limited to the 3' mRNA, yet with the use of different anchoring enzymes and sequencing efficiency, a higher understanding of expression dynamics can be envisaged.

The very nature of a SAGE library facilitates its comparison to other libraries, whether those libraries are derived from the same cells in experimental systems, cells from related lineages or even cells from unrelated lineages. The generation of SAGE libraries from different source material has potential to accelerate the understanding of cellular phenotype as well as investigations into restricted expression, thus bridging the gap between genotype and phenotype.

In summary, the transcriptome of NHMCs is well represented by genes of the cytoskeleton, ribosomal proteins and transcription factors. Also well represented was the basement membrane specific type IV collagen and other components associated with the mECM. Contractile apparatus were present at high abundance, many as 'smooth muscle' isoforms. Taken together these results support the functional role of the MC as a producer of the mECM and association with the GBM, affecting a contractile function. These properties belie the myoblast lineage of the MC.

SAGE is sensitive to variation in transcriptional products and was able to support the transcription of two genes from the RTN4 locus. Further, SAGE predicted peculiar transcription of CTGF with the generation of and anti-sense tag.

The digital nature of SAGE libraries facilitates the *in silico* comparisons of any SAGE library. Comparing SAGE libraries gives insights into the restricted expression of genes and demonstrates that SAGE can be used as a system of classification, which can be conducted at the library level (using tags, which are sensitive to 3' transcriptional variation) or at the transcriptome level (using genes). Either way SAGE can link the genotype to phenotype at a level not previously achieved.

CHAPTER 7

7 ANALYSIS OF DIFFERENTIAL TRANSCRIPTION

7.1 INTRODUCTION

Understanding the way a cell responds to stimuli in a disease model facilitates the identification of therapeutic targets. Stimuli initiating transcription can be monitored by sampling transcriptomes from normal and experimental models. Experiments in CHAPTERS 4, 5 & 6 demonstrated that the generation and properties of the NHMC SAGE libraries were stable and faithfully represented the transcriptome of NHMCs. In work described in this chapter, the NHMC SAGE libraries were separated into those generated from cultures grown under physiological concentrations of glucose and high concentrations of glucose, which mimic a diabetic environment, and the transcriptomes were compared in an attempt to identify glucose responsive genes.

Preliminary experiments defined the model and determined that it was responding to glucose by increasing transcription of TGFβ1, the classic marker of DN. Two SAGE libraries were generated, one from normal glucose cultures (LG) and one from high glucose cultures (HG). The experiments in this chapter describe the differential analysis of candidate genes in three stages. First, candidates representative of abundance classes and previously described genes were selected to assess the reproducibility of the culture protocol and RT-PCR technique. Next, genes identified during the preliminary GeneFilter analysis were analysed to assess the reliability those experiments. Finally, the SAGE derived transcriptomes were compared and candidates selected to assess the reliability of SAGE to identify differentially transcribed genes. In each stage, four experimental points were used. Proliferating cells 'P', and cells grown for 96h in normal glucose 'L', high glucose 'H' and equimolar mannitol 'M'.

In addition to transcription in NHMC cultures, a transformed MC, HMCL, was also analysed for a selection of genes. This directly compared two widely utilised culture models and assessed their similarity under the same conditions. The use of global transcription technology presents important issues regarding the phenotypes of cells. Being able to interrogate the transcriptome of cells at such high resolution will reveal differences in cells that had previously been assumed similar enough for substitution in model systems. Immortalised human mesangial cells (HMCL) are widely used and accepted as a convenient substitute for primary mesangial cells (Sraer et al., '96).

7.2 REAL TIME RT-PCR ANALYSIS

7.2.1 TRACKED CANDIDATE GENES

To begin, a group of genes were selected based on previous reports of altered transcription in similar culture models. Although they were identified not to alter in SAGE they were tested using rtRT-PCR to either confirm SAGE or the culture protocol. The validity of the NHMC SAGE libraries to detect and quantify genes was confirmed in CHAPTER 4 with a high degree of reliability, as assessed by correlation analyses, and so there was confidence that the data accurately reflects the transcriptome. However, there did appear a degree of uncertainty when comparing the abundances according to SAGE and rtRT-PCR, as assessed by rtRT-PCR and the SAGE determined abundance (see CHAPTER 5). Many of the genes tracked by their tags in the SAGE library showed minimal changes in abundance and so suggested no differential transcription. Yet, in preliminary experiments TGFβ1 transcription did show an increase in transcription in response to glucose. This is the classic indicator of DN and so the culture was considered valid.

To test the SAGE library for its ability to quantify differential gene transcription, a panel of genes was chosen that had either previously been described to respond to glucose or were representatives of abundance classes. Examples included high abundance genes like FN-1, TAGLN and collagen isoforms as well as medium to low abundance genes like TGFβ1 and Activin A (see TABLE 7.1). The objective of these experiments was to determine the reliability of SAGE to detect differences in transcription.

These 'tracked candidates' were segregated into arbitrary groups for analysis. In Set1, (FIGURE 7.1) all genes strongly increased transcription in response to the removal of serum and the extended quiescence (P compared to L, H and M). More subtle changes were noted between L cultures compared to H and M culture with increases in A2SMA (+1.2), COLL Ia2 (+1.5) and COLL IIIa1 (+1.1) being statistically significant ($p< 0.05$). These increases were also seen in the osmotic control (L v M) and so may be a response to osmotic stress rather than glucose.

TRACKED CANDIDATES

Tag Sequence	LG	HG	Fold Change		UniGene ID (Hs.)	Gene	References
			SAGE	Reported			
AAGATCAAGA	207	238	1.1	2.2 late	195851	α2SMA	(Sanai et al., '00)
ATCGTGGAGG	26	17	-1.5	-	727	ACTA	-
GATGAGGAGA	35	32	-1.1	Presence	179573	COLLIα2	(Wogensen et al., '99)
CCACAGGGGA	16	21	1.3	-	119571	COLLIIIα1	-
GACCGCAGGA	129	119	-1.1	1.4-1.5	119129	COLLIVα1	(Ayo et al., '90)
GAGCCTGGAT	27	15	-1.8	-	9004	CSPG4	-
TTTGCACCTT	131	129	-1	3	75511	CTGF	(Murphy et al., '99)
CTTTGAACGA	19	18	-1.1	-	75511	CTGF.2	-
TGCAATATGC	30	19	-1.6	-	750	FBN1	-
TGCATCTGGT	44	35	-1.3	Induced by glucose removal	75410	HSP70	Lee et al., '83
ATGTGAAGAG	226	198	-1.1	-	111779	SPARC	-
ACAGGCTACG	282	280	-1	-	75777	TAGLN	-
GGGGCTGTAT	1	3	3	2	1103	TGFβ1	(Yamamoto et al., '93) (Nakamura et al., '93)
ATCTTGTTAC	120	135	1.1	1.4-1.5	287820	FN1	(Ayo et al., '90)

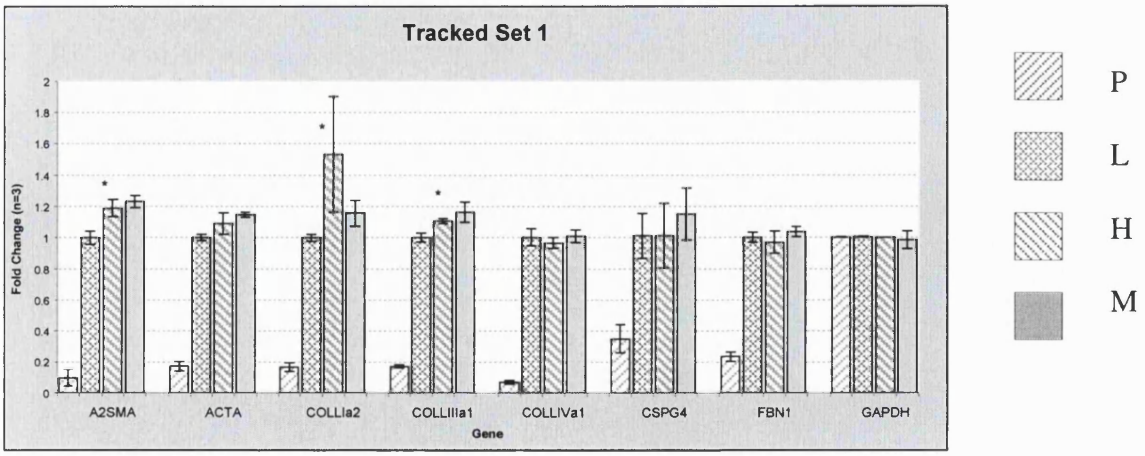
TABLE 7.1. SUMMARY OF TRACKED CANDIDATES.

This selection of genes were chosen based on their representation of different abundance classes or reported differential transcription in the model system (Description of gene abbreviations can be found in Appendix 1). Two primer sets were created for CTGF (CTGF and CTGF.2) in an attempt to assess the peculiar tag distribution seen in the SAGE library. Differences in the SAGE and reported change in transcription may be in part due to the differences in culturing conditions

Gene SET1	Fold Change Experimental Points			
	P	L	H	M
A2SMA	0.01	1	1.19*	1.23
ACTA	0.17	1	1.09	1.15
COLLIa2	0.17	1	1.53*	1.16
COLLIIIa1	0.17	1	1.11*	1.16
COLLIVa1	0.07	1	0.97	1
CSPG4	0.35	1.01	1.01	1.15
FBN1	0.23	1	0.97	1.04

FIGURE 7.1. GRAPHS AND TABLE OF PCR DATA 'TRACKED SET 1'.

Illustrating the difference in expression as determined by real time RT-PCR in the sequence P, L, H, M. Induction and repression data are expressed as average fold change from triplicate analysis of three independent cultures. (* $p<0.05$, L compared to H)



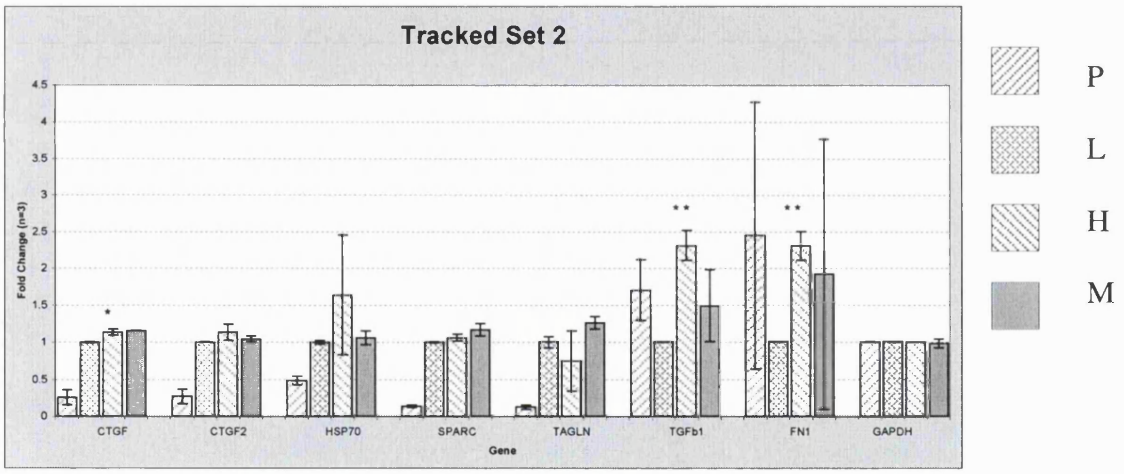
In contrast to Set1, in the second set some genes did not increase transcription in response to serum removal (FIGURE 7.2). CTGF/CTGF2, HSP70, SPARC and TAGLN

were highly up regulated upon the removal of serum but TGFβ1 and FN1 appeared down regulated when the serum was removed. Further, CTGF, TGFβ1 and FN1 showed significant up regulation in response to glucose (L compared to H; +1.15 $p< 0.05$ for CTGF and +2.3 $p<0.01$ for TGFβ1 and FN1). Again, apart from TGFβ1 these increases were also seen in the osmotic control and thus not attributable to glucose alone.

Gene SET2	Fold Change			
	Experimental Points			
	P	L	H	M
CTGF	0.26	1	1.14*	1.16
CTGF2	0.27	1	1.13	1.04
HSP70	0.49	1	1.65	1.06
SPARC	0.14	1	1.06	1.17
TAGLN	0.13	1	0.75	1.27
TGFb1	1.71	1	2.30**	1.5
FN1	2.46	1	2.28**	1.93

FIGURE 7.2. GRAPH AND TABLE OF PCR DATA 'TRACKED SET2'.

Illustrating the difference in expression as determined by real time RT-PCR in the sequence P, L, H, M. Induction and repression data are expressed as average fold change from triplicate analysis of three independent cultures. (* $p<0.05$, ** $p<0.01$, L compared to H). Modulation of CTGF was considered significant in these experiments.



With the exception of TGFβ1, all the genes that displayed increased transcription in response to glucose showed corresponding increases in response to mannitol ($p<0.05$ H compared to M). Again, this suggests that any change in transcription in these genes may be due more to osmotic stress.

Three observations can be clearly seen from these experiments. First, a change in transcription can be seen between proliferating cells 'P' and cells within the culture protocol 'L, H and M'. This indicates that alterations in transcription can clearly be detected. Second, the absence of altered transcription (in most of the candidates) in response to either glucose or mannitol suggests that in this culture protocol there is little effect on transcription of candidate genes of glucose alone. Finally, the increased

transcription of CTGF, A2SMA, TGFβ1 and FN-1 concurs with previous reports that these genes respond to high glucose or mannitol by increasing transcription (see TABLE 7.1). Interestingly types III & I collagens were also increased in response to glucose and mannitol although to a lesser extent.

7.2.2 SUMMARY OF TRACKED GENES

Most of these results were predicted from the SAGE analysis. The case for FN-1 is interesting. SAGE suggested that FN-1 transcription would remain essentially constant but there is a moderate up regulation in this system. However, the RT-PCR data is not consistent with the abundance of FN-1 relative to other genes. The normalised Ct values for this particular FN-1 amplicon (4.09) indicates that the FN-1 transcript is some 3-fold lower than say FBN1 (2.46) and some 6-fold lower than A2SMA (1.46). According to SAGE FN-1 is some 4-fold higher than FBN1 and 2-fold lower than A2SMA. Investigation to clarify the true abundance of FN-1 and the reliability of the RT-PCR and SAGE data regarding the changes in transcription is required

The classic transcriptional indicator of DN is TGFβ1 and increases in transcription of TGFβ1 in both DN and MC grown in high glucose are well documented. The gene also increases in expression in this culture system according to SAGE and rtRT-PCR and independent of the osmotic control culture (equimolar mannitol). The other genes in this panel of candidates were predicted, according to SAGE, not to alter substantially. Apart from FN-1, this was the case. Minor increases were seen in A2SMA, COLL Iα2 and CTGF.

Taken together this demonstrates that the culture system is responding in a reproducible and predictable way. This suggests that if tracked genes other than TGFβ1 are changing transcription in response to glucose then this is occurring after the time scale used in these experiments.

7.3 CANDIDATE GENES DETERMINED FROM SAGE ANALYSIS

7.3.1 PRIMARY COMPARISON

Matched pairs of tags were compared directly from the SAGE report output prior to mapping, as an initial measure of difference would be useful in describing the transcriptomes. Efficient analysis required the use of filters such as the statistical analysis and 11th bp resolution, but the initial measure of differential tag abundance was compiled based on a +/- 5-fold abundance. This would more than likely contain many false candidates, but a precedent of three-fold was suggested by key reports that sampled libraries to this level (Chen et al., '98a). The SAGE libraries in these reports were generated from extremely dynamic models of primary mast cell stimulation with LPS. Such potent gene activation would stimulate a large variation between key genes involved in mast cell activation. As the high glucose and low glucose NHMC transcriptomes appeared more similar, suggesting glucose stress on NHMCs maybe more chronic, this primary filter was increased from 3-fold to 5-fold, in an attempt to favour detection of positive candidates. From this initial list, 'reliable' tags were selected using filters described below.

A primary comparison was conducted on the raw output form the SAGE analysis. A total of 23,001 tags from the LG library and 23,128 tags from the HG library were used in a primary comparison to identify differentially transcribed tags. Because the 2° transcriptomes contain the most useful data regarding gene abundance but does not contain any redundant tags, which may cause a loss of data, those tags that mapped to more than 3 genes were removed from the analysis and tags that mapped to 2-3 genes were resolved using the 11th bp (see below). These abridged transcriptomes formed the primary comparison (see APPENDIX 5 for the full list of differential tags and genes to which they map).

The transcriptomes appeared remarkably similar. Of the 4553 matched pairs, 26 tags were increased more than 5-fold and 133 tags were decreased more than 5-fold with the majority changing little. This concurs with the Gene filter analysis in CHAPTER 3, which predicted few changes in the transcriptional patterns.

7.3.2 DETERMINING STATISTICAL SIGNIFICANCE TO CHANGES IN TAG FREQUENCY

Simple statistical tests such as those used to determine the relationship between each transcriptome in CHAPTER 4 are less informative when analysing matched pairs from random samples of complex populations. A Bayesian analysis has been used in many SAGE projects because of the need to be able to infer statistical probability from an empirically determined prior assumption. As discussed in CHAPTER 2.8.3, the Bayesian analysis takes the form of a Beta distribution that has been estimated from a proportional distribution. Briefly, a 'prior distribution' is determined by the distribution of the proportion of tags in matched pairs from two populations, which can be plotted as a probability density function in the range [0,1]. This density function can be described by a Beta function with beta parameters 'a' and 'b' (see EQUATION 2.2) so a mathematical derived definition of prior probability can be assigned to each matched pair and solved for the 'posterior probability' 'p' after observing sampled frequency. Each SAGE analysis requires the estimation of these 'prior assumptions', which are obtained from the empirical data and is particular for each comparison. A convention of $p > 0.5$ has been adopted to filter matched tag pairs for further analysis. This translates to a 50:50 chance of the matched pair being increased, or decreased by a defined level.

The distribution for proportion of tag frequency in each sampled population was plotted in the range [0,1] (FIGURE 7.3). As expected, the distribution is not a linear, or uninformative, indicated by equal proportional distribution of all tags, but approximately symmetrical around $x = 0.5$ indicating that the majority of tags do not change. This agrees with many other analyses of SAGE libraries and infers the Beta function is also symmetrical with parameters equal. Determining the parameters requires the estimation of which Beta function best fits the data based on the observed data and a level confidence in the data. The higher the beta parameters then the higher the concentration 'x' around 0.5 and the more data required to infer a given change in tag frequency. Previous reports have suggested beta parameters of $a = b = 2$ (Chen et al., '98a) but these will not reflect the data from this SAGE analysis. Plotting various Beta functions over the empirical data suggests that very stringent beta parameters will most accurately reflect the data (see FUGURE 7.3). Parameters of $6< (a = b) <8$ were most accurate in fitting the NHMC data and these were used to infer a probability of

differential tag abundance. The higher the confidence that the data changed between distributions then the lower the beta parameters used to estimate the prior function, whereas the higher the beta parameters then the less the confidence that the distributions change.

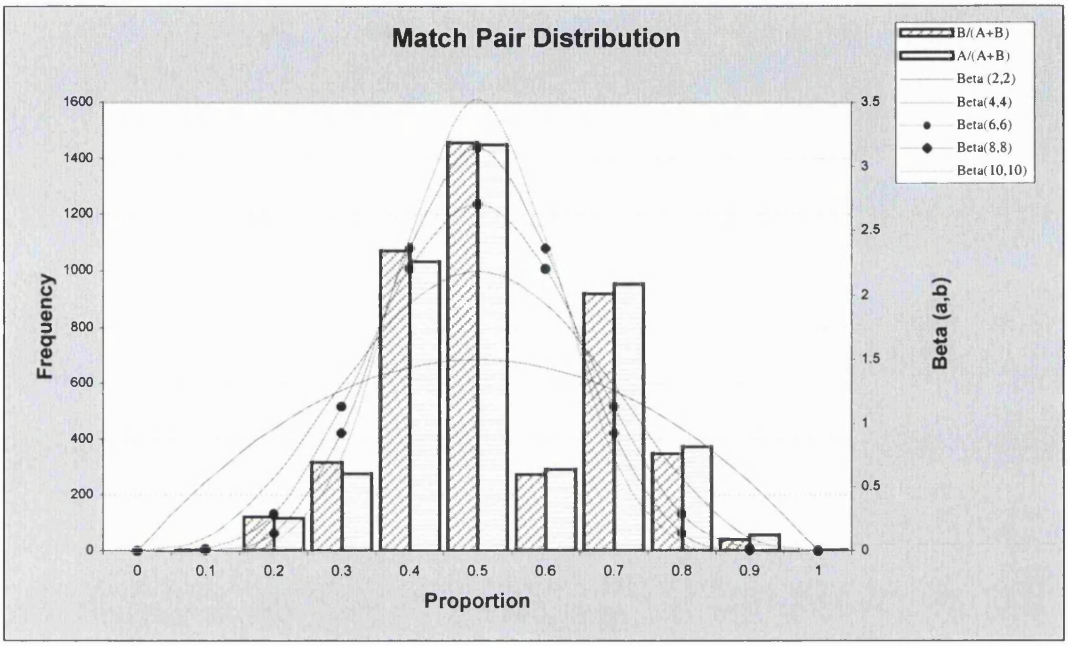


FIGURE 7.3. PROPORTIONAL DISTRIBUTION FOR ALL MATCHED TAG PAIRS A&B.

Tags present only once in the combined libraries were removed and the proportional representation of each pair plotted against frequency (left hand y-axis scale). Induction, (A/(A+B)) and repression, (B/(A+B)). Representative symmetrical beta functions were superimposed across the matched pair proportional distribution (right hand y-axis scale). From the graph it can be suggested that the most accurate proportional distribution will lie between Beta (6,6) and Beta(8,8). Greater confidence that the data changes little selects Beta (8,8) as the prior function, while if there were greater confident that the data changes between SAGE libraries, then Beta (6,6) is selected.

The prior pdf selected to estimate this data, inferred that a very high level of difference between matched tags was required before an acceptable statistical significance would be assigned. The 'posterior probability' that a tag is altered by at least 3-fold was calculated for each matched pair. In this SAGE analysis this resulted in no tags bearing a differential frequency of magnitude sufficient for justification of differential analysis (TABLE 7.2). Many reports describe the altered transcription of genes in MC grown under high glucose and so overwhelming evidence exists that this is the case. The fact that they cannot be resolved in this SAGE analysis does not demonstrate that these changes are false, simply that the libraries may not been sampled sufficiently to achieve statistical significance using these analytical techniques. Many of the genes present in the SAGE libraries and implicated MC glucose stress did not exhibit differential tag abundance but some did (e.g. TGFβ1). If this were the case, then

by relaxing the statistical parameters to favour the detection of differential patterns over stringent filtering of false differential data, it may be possible to discover new genes whose transcription is altered in this system. However, this may also introduce more positives that are false and so lead to inefficient selection.

7.3.3 SELECTION OF RELIABLE SAGE CANDIDATES

7.3.3.1 PRIMARY TAGS

All tags from the 1° list were subjected to a series of filters. First the mapping information was inspected and the tag was rejected if it was not the major tag for that gene in the library i.e. if the tag showing differential abundance was superseded by a more abundant tag that showed no change in abundance. This filter was relaxed if the tag under inspection contributed more than the estimated error of tag generation discussed in CHAPTER 4 (6% to the total abundance of the gene), which permitted alternative transcription products.

7.3.3.2 STABLE ACCUMULATION OF TAGS

The abundance of tags in each of the sub-libraries was also collated to examine the stable accumulation of tags across the sub-libraries. It was considered that a stable accumulation of tags would represent a 'more likely' difference between the two culture systems. The majority of tags appeared to accumulate at a stable level across the sub-libraries but because candidates were present at a relatively low abundance, this process was considered redundant (see below).

7.3.3.3 LEVEL OF SAMPLING

A further consideration was the level of sampling. The beta parameters used to infer differential abundance are really only useful with matched pairs for which there is little empirical data. As the sampling level increased (e.g. for genes of high abundance) the parameters becomes less useful. With this in mind, the levels for tags required to imply a change were determined by solving the Beta function for a variety of hypothetical tag pairs. Thus, p (fold change > ±3) would be 0.51 for tag pairs 5:27, 10:42, 30:102, 40:132, and 50:162. Similarly, p (fold change > ±2) would be 0.5 for tag

pairs 5:15, 10:25, 30:64, 40:83, 50:103. Inspecting the candidates revealed that none of the high abundance tag pairs fulfilled these levels.

PRIMARY SAGE CANDIDATES							
Tag Seq	11th Base	LG	HG	Change	Beta (0.75,6,6)	UniGene ID	Gene
TAAAATGAAA		0	9	9	0.38	Hs.24950	Regulator of G-protein5
GGCTGGTCTG		17	3	-5.67	0.37	Hs.50724	OVARC1000640
AAAGTTCGTA		1	10	10	0.3	Hs.82306	Destrin
ATACAAGAGC		1	10	10	0.3	No Match	
TCACCCACAC	C	6	23	3.83	0.28	Hs.234518	RP L23
TCACCCACAC	C	6	23	3.83	0.28	Hs.322680	EST
GTGCTGAAGG		9	1	-9	0.26	No Match	
AAGATTTTAG		0	6	6	0.23	Hs.21537	Protein phosphatase 1β
CTCATCAGCT		0	6	6	0.23	Hs.104125	Adenylyl cyclase-associated protein
CTTCAGCTAA		0	6	6	0.23	No Match	
GACTGTTAAT		0	6	6	0.23	Hs.118684	Stromal cell-derived factor 2
TTATGTATCA		0	6	6	0.23	Hs.169300	TGFβ2
GACTCACTTT		2	11	5.5	0.23	Hs.699	Cyclophilin B
GCACCTTATT		2	11	5.5	0.23	Hs.125078	Ornithine decarboxylase antizyme 1
GCGACAGCTC		6	0	-6	0.23	Hs.184582	RPL24
GGCCAAAGGC	C	6	0	-6	0.23	Hs.213701	EST
ACCATCAATA		1	8	8	0.21	Hs.169476	GAPDH
CAATGTGTTA		1	8	8	0.21	Hs.74823	NADH-DH1 α 1
TAAAAGACAA		1	8	8	0.21	Hs.77196	Spectrin, alpha
TCTCAATTCT	T	1	8	8	0.21	Hs.146409	CDC 42

TABLE 7.2. PRIMARY SAGE DETERMINED TAGS OF DIFFERENTIAL FREQUENCY.

The probability '*p*' that they are real and at least a three-fold change is determined by a Beta (0.75,6,6) analysis. All tags suggest $p < 0.5$ and thus little significance. Changing the beta evaluation to detect changes of at least two fold (Beta (0.66,6,6)) the probability that there are differences for all tags rise above 0.5. 11^{th} base evaluations are noted where significant. Despite that apparent high probability that all of these genes are increased at least two fold, most of these tags are redundant for other genes and are removed from the analysis.

7.3.3.4 RESOLVING THE 11^{TH} BASE PAIR

It was also considered that some tags might be excluded from the analysis due to their redundancy with other genes. Resolving tags that mapped to more than 10 genes was considered impractical for this analysis, but an attempt to resolve tag mapping was made for selection of the candidate tags that each mapped to at most 5 genes. Of the 100 tags selected, 48 tags resolved using the 11^{th} base but no tags were used for further differential analysis as this resolution reduced their differential tag levels below the acceptable level (see APPENDIX 5).

An example of this 11th bp resolution is as follows. The tag GCTTTCCATCT (1/6) (n=1 for LG, n=6 for HG) maps to two genes, HLA-B associated transcript (Hs.55296) and an EST cluster associated with tumour progression (Hs.78768). Inspecting representative sequences from these UniGene clusters revealed that the 11th base for Hs.55296 to be 'T' and for Hs.78768 is 'G'. When the extended tags were identified in the transcriptome the respective abundance was altered to GCTTTCCATCTT (1/4) and GCTTTCCATCTG (0/2). In comparison to this devaluation of tag frequency some genes could be excluded from the redundancy with the 11th bp, e.g. AGGACAGAAG (1/4) maps to Hs.183698 (RP L29) and Hs.198625 (MMP25), however, resolving all the tags in the SAGE library determined the 11th bp to be 'G', which maps only to Hs.198625 (MMP25).

SAGE CANDIDATES

Tag Sequence	LG	HG	Fold	UniGene ID (Hs.)	Gene Symbol	Gene	Reference sequence
GGTGAGACAC	3	1	-3	164280	ADPtr	ADP/ATP Translocator	J03592
ATCCGTGCCC	4	1	-4	141011	CALM3	Calmodulin 3	NM005184
TGCCTCTGCG	70	66	-1.1	75564	CD151	CD151 antigen	NM004357
CACCCCTGAT	0	4	4	173724	CKB	Creatine Kinase B	NM001823
GACCAGAAAA	4	0	-4	180714	COX6a1	Cyc Oxidase 6a1	NM004373
CCGTGCTCAT	4	1	-4	9857	CR	Carbonyl Reductase	NM016286
ATGAGCTGAC	1	7	7	695	CSTB	Cystatin B	NM000100
CCATTTTCTG	0	5	5	198899	DCIP-1	D-type cyclin interacting protein	NM012142
AAAAAACCCA	4	1	-4	111680	ENSa	Endosufine alpha	NM004436
TACCATCAAT	80	77	-1	169476	GAPDH	Glyceraldehye 3DH	NM002046
CCAACCGTGC	7	2	-3.5	75207	GLO1	Glyoxalase 1	NM006708
GAAATTTAAA	1	4	4	274472	HMG1	High Mobility Group 1	NM002128
CAATGTGTTA	1	8	8	74823	NDUFa1	NADH dehydrogenase1a1	NM004541
ACTACTAAGG	8	5	-1.6	2820	OXYrc	Oxytocin Receptor	NM000916
TAACCCAACA	5	0	-5	1869	PGM1	PhosphoGlucoMutase 1	NM002633
ATGATGCGGT	6	1	-6	41072	SERPIN b6	Serine protease Inhibitor type b6	NM004568
TTATGTATCA	0	7	7	169300	TGFβ2	TGF β2	NM003238
CTCTTCGAGA	8	2	-4	76686	GPX1	Glutathione peroxidase 1	NM000581
AAATAAAGAA	4	1	-4	790	mGSH	Microsomal GSH transferase 1	BC005923

TABLE 7.3. SUMMARY OF RELIABLE SAGE DIFFERENTIAL CANDIDATES.

Reference sequences were used to design primers for this selection of genes. The genes were chosen as representative of differentially transcribed genes according to SAGE analysis

7.3.3.5 FINAL LIST

A complete list of all candidates selected based on these conditions is presented in APPENDIX 5. For further analysis a selection of genes from this list were used to test

the SAGE-observed compared to actual differential transcription (TABLE 7.3). These candidates were expected to be the most reliable differential genes. TGFβ1 and FN-1 were included as positive controls and GAPDH and CD151 was included as a normalising factor.

7.3.4 REAL TIME RT-PCR ANALYSIS TO TEST CHANGES IN CANDIDATES

Real-time RT-PCR was conducted in duplicate for three independent culture series at four points, proliferating cells (P), 96h in normal glucose (L), 96h in high glucose (H) and 96h in equimolar mannitol (M). The results were normalised to GAPDH, as this gene does not alter in the SAGE library, and then calibrated to the normal glucose culture 'L'. Normalisation to GAPDH ensured that differences seen in SAGE could be compared to the rt-RT-PCR. As can be seen from the graphical representation (FIGURE 7.4a,b & c) all of the candidate genes showed insignificant changes in transcription across the three experimental points, L, H and M. As expected there is a general induction of transcription between the proliferating cells, P, and the serum starved experimental points with many of these differences being statistically significant ($p<0.05$).

7.3.5 SUMMARY ON THE ANALYSIS OF SAGE DETERMINED CANDIDATES

Filtering criteria were used to reduce the primary list, which contained many false positive data, to a list of reliable candidates. Redundant tags were removed where the 2° tag level was below the calculated level of contamination (6%). Tag frequencies were examined for stable accumulation across sub-libraries but candidate sampling was low. Where possible, conflicts between high abundance and redundant tags were resolved by extending the tag to the 11^{th} base. A selection of candidates was used to assess both the differential transcription of genes in response to glucose and the efficiency of the Beta analysis and filtering criteria on the primary comparison.

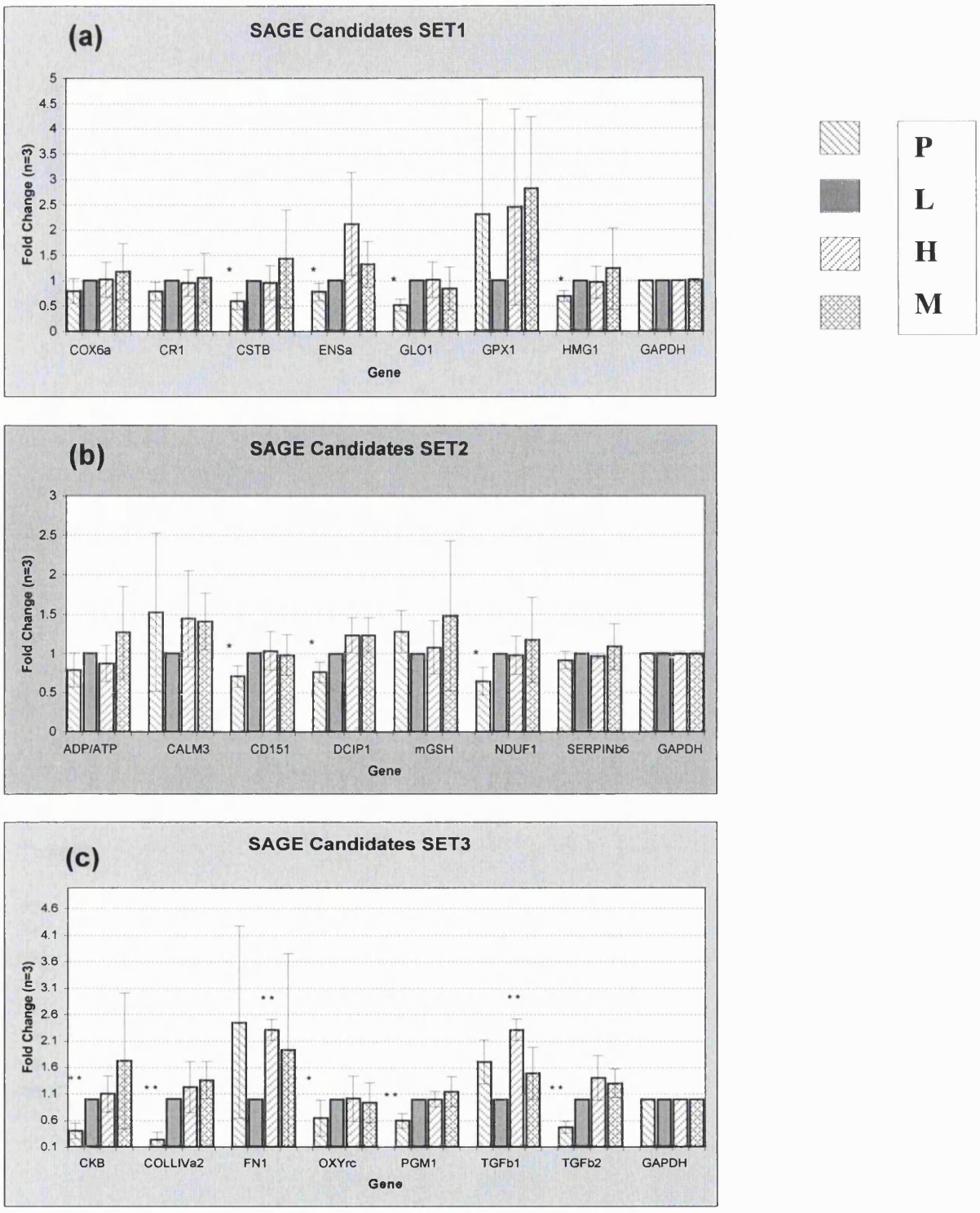


FIGURE 7.4 A,B,C. GENES PREDICTED TO ALTER TRANSCRIPTION AS DETERMINED BY SAGE.

Genes are grouped in sets of four corresponding to four experimental points P, L, H and M. As with the tracked genes in the previous section the greatest changes in transcription were seen between proliferating cells 'P' and the experimental samples L, H and M and the only significant changes seen between L and H was for TGFβ1 and FN-1 (**p<0.01, *p<0.05)

In three independent cultures many of the candidates changed transcription in response to the removal of serum (P compared to L, H and M). For the majority of these genes this was a strong up regulation but for several was measured to be a down-regulation. When comparing normal glucose 'L' to high glucose 'H' only two of the candidates suggested an increase in expression, TGFβ1 and FN-1 as assessed by rtRT-PCR but only TGFβ1 was predicted to increase in abundance based on SAGE and this prediction was excluded from the selection procedure due to low sampling level. From these experiments, it can be concluded that firstly for all but one gene the RT-PCR analysis concurred with the Beta analysis of the SAGE libraries that there was no significant difference between the two populations. This suggests that either glucose has very little effect on the NHMC or that the transcriptomes of the two culture systems have not been sufficiently sampled for accurate predictions to be made. Given that the gene shown to respond to glucose using RT-PCR (TGFβ1) was present in the SAGE library but excluded from candidature due to low tag frequency, and that there is much literature describing real effects of glucose on MC, it is likely that the NHMC transcriptomes have not been sampled sufficiently in these libraries.

7.4 CANDIDATE GENES FROM GENEFILTER ANALYSIS

7.4.1 PRIMARY COMPARISON OF GENEFILTER SIGNALS

The preliminary experiments described in CHAPTER 3 identified potential differential data based on hybridisation. In the following experiments these GeneFilter candidates were tested more rigorously using the same rtRT-PCR protocols for the previous sections.

The GeneFilter GF200 is claimed to contain targets for 5000 genes. Some of these genes will be housekeeping genes used for normalisation algorithms in the analysis software and some are positive and negative controls. Of these potential targets some 937 signals were measured to be above the limit of resolution for this technology (see APPENDIX 6). Of these, some 15 genes were persistently altered in signal intensity (see TABLE 3.6). As with convention, a 'greater than 2 fold' threshold

was applied to these candidates and thus 8 genes were selected to test the reliability of the gene filter (see TABLE 7.4).

There were a high proportion of uncharacterised genes. Of the eight candidates persistently up-regulated 5 were ESTs, possibly indicating important novel candidates for further study. Of the three characterised genes identified in the gene filter analysis each was connected with a different biochemical pathway. CD58 is a counter receptor and member of the immunoglobulin super family. Binding between CD2 and CD58 optimises immune recognition and facilitates the interaction between T-cells and antigen presenting cells (Sewell et al., '88, Barbosa et al., '86). SEC22A is a member of the SEC22 vesicle trafficking protein family and is believed to be involved in the early stages of the secretory pathway (Tang et al., '98). Finally, APPBP2 is a binding protein of amyloid precursor protein, a peptide baso-lateral sorting signal. These proteins are involved in the control and maintenance of cellular polarity (Zheng et al., '98). The presence of these genes in the SAGE library was investigated. Both CD58 and APPBP2 were present but each at such a low level as to be uninformative regarding differential transcription.

GENE FILTER CANDIDATES						
Tag Sequence	LG	HG	GF Fold	UniGene ID (Hs.)	Gene Symbol	Gene
ATTTGCACCT	0	2	-2.1	84084	APPBP2	Amyloid beta precursor protein binding protein 2
TATTGTGCTG	1	1	3.2	75626	CD58	Lymphocyte function-associated antigen 3
-	-	-	-3.1	42029	EST3	-
-	-	-	2.3	326725	EST4	-
-	-	-	7.2	12097	EST5	-
-	-	-	2.1	183655	SEC22A	Sec22 homolog
-	-	-	2.2	233634	EST2	-
-	-	-	2	278634	EST1	-

TABLE 7.4. SUMMARY OF CANDIDATE GENES AS DETERMINED FROM GENEFILTER ANALYSIS.

Fold change determined from Pathways™ software and the presence of each gene in the SAGE analysis is indicated under LG and HG.

7.4.2 RT-PCR ANALYSIS OF GENE FILTER CANDIDATES

Each of these 8 genes was used in real time RT-PCR experiments to test the reliability of the GeneFilter analysis. Graphical representation is illustrated in FIGURE

7.5. As with all other candidate analyses in this set of experiments there appeared a marked change in transcription between proliferating cells and the experimental samples. This change was both an induction (SEC22A, APPBP2, CD58 & ESTs 1,3,4) and suppression of transcription (ESTs 2 & 5). When high glucose cultures were compared to low glucose cultures only APPBP2 showed moderate but significant suppression of transcription (-1.3fold, L compared to H, $p<0.02$) together with no suppression in the corresponding osmotic control 'M'. EST3 displayed increased expression (+1.5, L compared to H, $p<0.05$) but this was contrary to predictions from GF analysis. All other significant changes were seen between proliferating cells 'P' and the experimental points L, H and M. CD58 suggested an up-regulation of transcription but sample-to-sample variation reduced the significance.

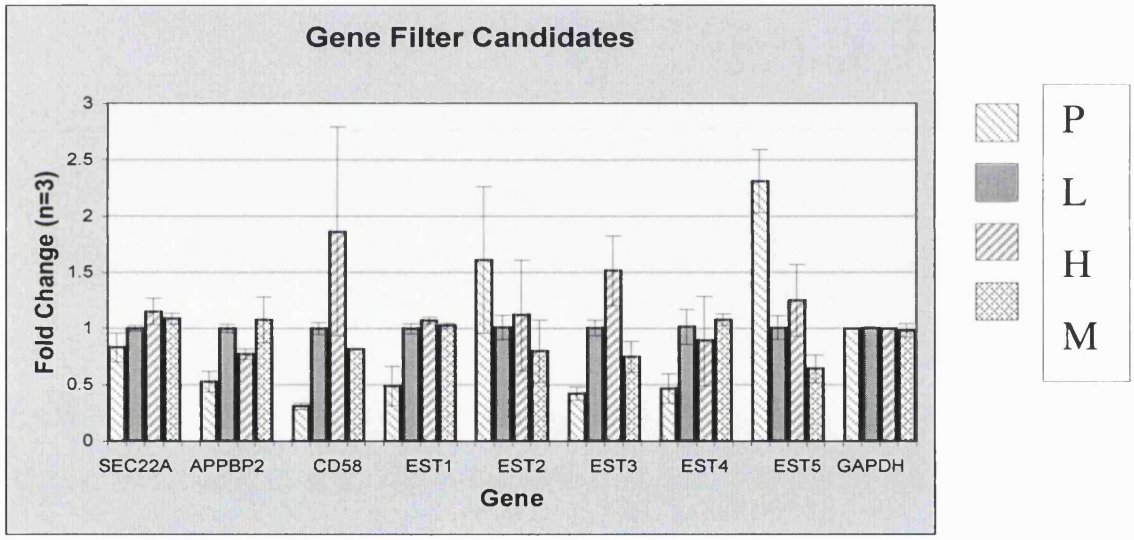


FIGURE 7.5. GRAPHICAL REPRESENTATION OF THE RT-PCR ANALYSIS OF GENE FILTER CANDIDATES.

All data is normalised to GAPDH and calibrated to L. Generally the results show that while there are large differences between proliferating cells, P and the experimental sample L, H & M, there is little change in transcription for all but two of the genes with L compared to H. Differential transcription of APPBP2 concurs with GF analysis while that of EST3 is contrary to GF analysis. Both differential levels were low (APPBP2, -1.3 $p<0.02$, and EST3, +1.5 $p<0.05$).

7.4.3 SUMMARY OF THE GENEFILTER ANALYSIS

The gene filter analysis proved unreliable with regard to differential transcription analysis with only one gene confirmed down regulated in response to glucose stress and one gene determined to be contrary to GF prediction. The GF

analysis did concur with SAGE and RT-PCR analysis, in that the cultures changed little in response to glucose.

7.5 RT-PCR ANALYSIS OF NHMC & HMCL

From personal observation using HMCL and this culture protocol, cells proliferate on a culture dish to confluence and then peel off the dish if not made quiescent by the removal of serum. NHMC will proliferate until confluent, when proliferation slows and they begin to form focal 3D structures (hillocks), while remaining firmly attached to the culture dish, even in the presence of 10% FCS. It has been postulated that because the ECM is linked to cellular phenotype and proliferation, then the immortalisation of the MCs has an effect on ECM components (Anderson et al., '94, Brady et al., '00). To test this a selection of ECM and cytoskeleton genes was compared between NHMC and HMCL. In this particular model of DN, the prominent use of HMCL has facilitated clarification of the mechanisms involved in glucose stress on MC. However, with the discovery of transformation sensitive elements in this primary cell system (e.g. transgelin), it may be important to investigate the relative transcription of a selection of genes prominent in the NHMC transcriptome. This may reveal fundamental differences between the cell systems subjected to the same experimental protocol. Any such differences may have implications to the physiological significance of *in vitro* data.

7.5.1 CANDIDATES FOR COMPARING HMCL TO NHMC

Various collagen isoforms are described in MC (Simonson et al., '89, Mene et al., '89). The most commonly discussed are the basement membrane and mECM specific type IV isoforms, but also present in MC, in both independent reports and the transcriptome generated as part of this thesis, are the types I & III isoforms. Types I & III collagen are not generally described *in vivo* but their increase in abundance has been described *in vitro* (Wogensen et al., '99). Generally, this has been attributed to an activated phenotype adopted but the MC *in vitro* (Stephenson et al., '98).

The cytokines TGFβ1, TGFβ2 and the growth factor CTGF are also present in MC with TGFβ1 and CTGF being characterised to play important roles in glucose stress of MC. TGFβ2 is the second of three members of the TGFβ sub-group and is rarely considered when discussing TGFβ1, although its presence has been confirmed in this model. TGFβ2 has been implicated in the prevention of glial scarring (Logan et al., '99) and locking cells in the G1/S checkpoint of the cell cycle (Chen et al., '99). Anti- TGFβ antibody has been widely used to demonstrate the requirement of TGFβ1 in DN, but many of the early antibodies had cross reactivity to other TGFβ family members (Mozes et al., '99, Ziyadeh et al., '00).

The matricellular SPARC has been described in MC and is involved is a variety of cellular processes including the induction of TGFβ1 (Bassuk et al., '00). The transformation sensitive and smooth muscle specific TAGLN, also was highly represented in the NHMC transcriptome (discussed in CHAPTER 6). Strong down regulation of TAGLN has been reeported in transformed systems and it is predicted that these experiments will also show this. Fibronectin was also included in this analysis.

7.5.2 RT-PCR ANALYSIS OF HMCL IN GLUCOSE STRESS

Initially, the effect of glucose on HMCL was assessed for TGFβ1 and CTGF, which have been reported to increase in transcription in response to glucose. Also included, due to its prominence in the NHMC transcriptome, was the related TGFβ2. The results of these RT-PCR experiments are graphically represented in FIGURE 7.6.

As can be clearly seen in FIGURE 7.6 both CTGF and TGFβ1 show significant up regulation in response to glucose when compared to both normal glucose and the osmotic control (CTGF +1.5 $p<0.05$, TGFβ1 1.6 $p<0.01$). This indicates that the culture system is responding in a predictable and reproducible manner. The change in TGFβ2 was not significant.

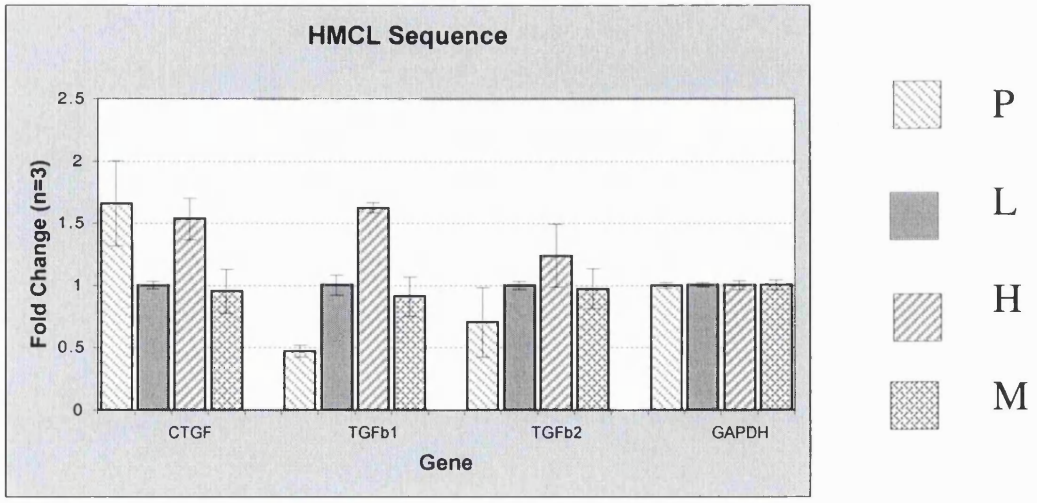


FIGURE 7.6. HMCL RESPONSE TO HIGH GLUCOSE PROTOCOL.

HMCL cells grown under high glucose were grouped for each of three genes to compare how these genes responded to glucose in this culture system. The sequence of each data bar as are for previous grouped experiments P, L, H, M. Errors are estimated with the 1st standard deviation. Clear induction of both CTGF and TGFβ1 is seen in response to glucose and not reflected in the osmotic control. TGFβ2 transcription did not alter significantly.

7.5.3 COMPARISON OF NHMC TO HMCL

Data from RT-PCR experiments demonstrate a clear difference in the expression of important ECM and cytoskeleton associated genes in the two cell types (TABLE 7.5 & FIGURE 7.7 a,b &c). The cytoskeleton gene TAGLN showed a strong suppression in HMCL compared to NHMC, consistent with the notion that TAGLN is a transformation sensitive gene (FIGURE 7.7a). The metabolic housekeeping gene GAPDH showed continued decrease in abundance across the time course but no difference between the two cell types. The matricellular SPARC and FN-1 genes were also highly repressed in transformed cells.

Interestingly, the type IV, III and I collagens were substantially different between cell types with HMCL expressing extremely low levels compared to the counterpart primary cultures (FIGURE 7.7c). While the presence of type IV collagen was used to characterise these cells, the absence of types I and III collagens were not (Sraer et al., '96). When compared to normal cells there was about 3 fold less COLL IV transcription in HMCL but there were three clear orders of magnitude less COLL I and COLL III in HMCL. Clearly, the transformation of MC has dramatic effects on both cellular proliferation and ECM production.

Transcription of TGFβ1 between HMCL and NHMC was comparable in magnitude (FIGURE 7.7b). Interestingly CTGF showed a large difference in abundance in transformed cells. This did not however change the response to glucose, which is calculated to be approximately +1.3-1.6 fold in the respective cell systems.

Gene	Fold Change	P (n=3)
COLLIα2	-3170.7	0
COLLIIIα1	-5513.5	0
COLLIVα1	-2.6	0
CTGF	-1.4	0.498
FN1	-42	0.001
GAPDH	0	0.869
SPARC	-1.7	0.041
TAGLN	-4.6	0.007
TGFβ1	-1.1	0.142
TGFβ2	-2.2	0.127

TABLE 7.5. DIFFERENTIAL GENE TRANSCRIPTION BETWEEN PROLIFERATING HMCL AND NHMC.

All values were normalised to GAPDH. Fold change is expressed as relative to counterpart cultures of primary NHMCs. Especially noted is the several orders of magnitude change in types I & III collagens compared to the relatively modest difference in type IV.

7.5.4 SUMMARY OF THE COMPARISON OF NHMC AND HMCL

Unlike type IV collagen, types I and III are not normally present in the mesangium or the glomerulus (Mene et al., '89, Davies, '94) and so their presence *in vitro* may indicate an activated state of the cells. This has been proposed for NHMCs that are starved of serum, where a strong up regulation of A2SMA (some 10fold in 24hr) is observed (Stephenson et al., '98) and FIGURE 7.1. The same may be concluded for SPARC, which is not generally present in the glomerulus but has a very high abundance in the NHMC transcriptome (Gilbert et al., '95, Pichler et al., '96). In addition, these experiments confirmed that TAGLN in NHMC is transformation sensitive and some 5-fold lower in transformed proliferating cells while some 50-fold lower in transformed quiescent cells.

Despite the severe differences in abundance of collagen isoforms the abundance of CTGF, TGFβ1 and TGFβ2 were present in HMCL at comparable levels to NHMC and very similar in their response to glucose, which concurs with previous reports on the use of HMCL (Sraer et al., '96, Wolf et al., '97, Wolf et al., '98). According to these experiments it would prove wise to examine the 'normal' expression of ECM components *in vivo* or *in vitro* with primary cells before extrapolating normal expression profiles from transformed counterparts.

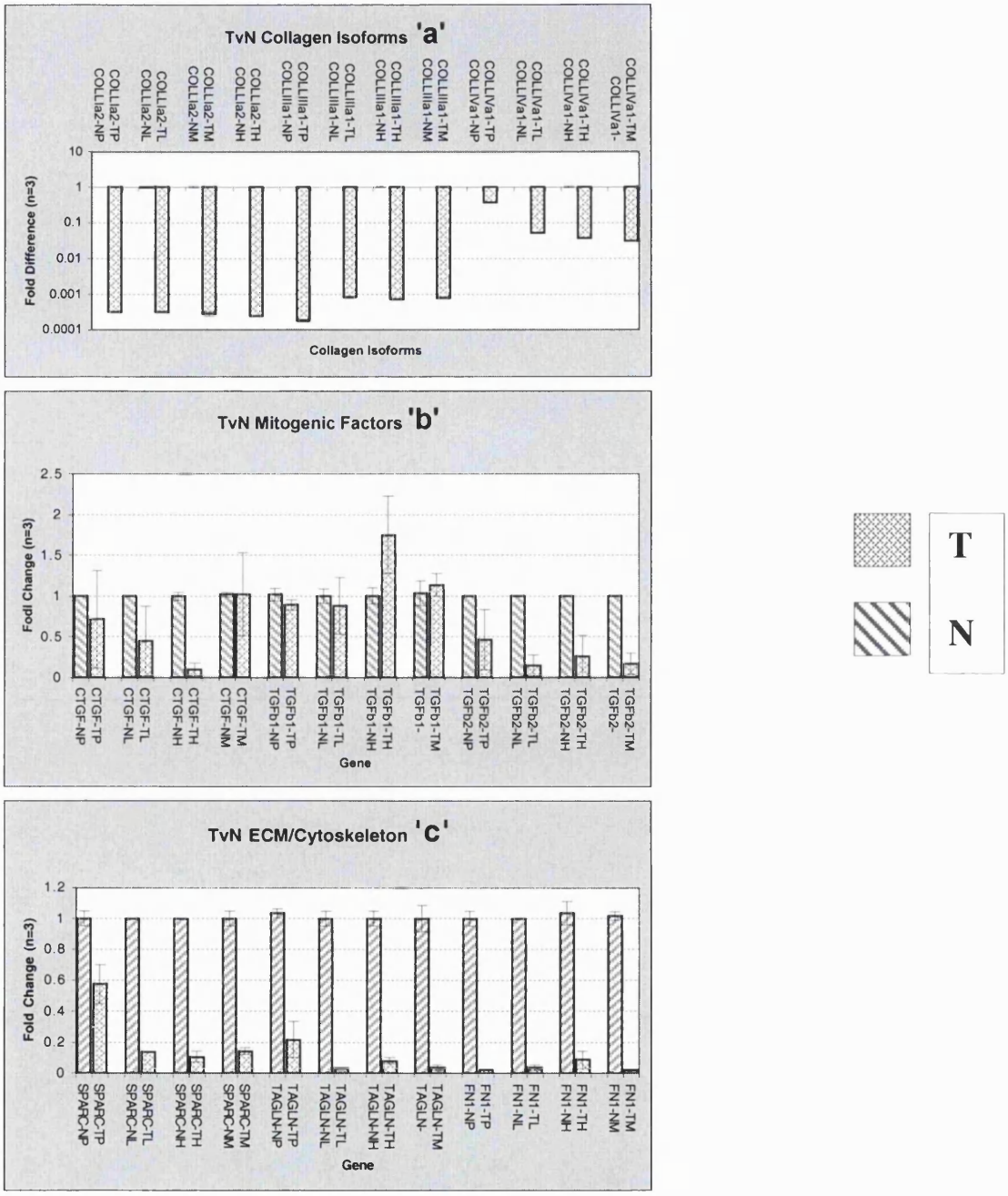


FIGURE 7.7 A,B,C. COMPARISON OF NHMC TO HMCL.

Graphical representations of HMCL compared to NHMC for collagen isoforms 'a', mitogenic factors 'b' and ECM/Cytoskeleton genes 'c'. Matched pairs for each of Transformed (T) and Normal primary (N) were normalised to GAPDH and calibrated to corresponding normal cultures. Estimation of errors uses the 1st standard deviation.

7.6 DISCUSSION

The culture series was tested for differential transcription using a set of genes known to alter transcription in high glucose, or representatives of abundance in the NHMC transcriptome. Analysis of candidate genes in the model demonstrated that while the classic indicator of glucose stress in MC was present and responding in a predictable manner all other candidates were unaffected by high glucose alone with some appearing to respond to a general osmotic stress (FN-1, CTGF). Many similar models subject MC to periods of glucose stress well past the 96 hours used in this model (Nakamura et al., '93, Yamamoto et al., '93, Murphy et al., '99). It is possible that the other changes seen in this model, such as a general increase in ECM proteins, occurs outside the time scale of this analysis. However, while it is possible to redesign the culture protocol to extend the glucose stress the object of these experiments was to evaluate the differential transcription specifically at 96h, as this was where the SAGE libraries had been constructed. As TGFβ1 was shown to increase at this time point, no further experiments were undertaken.

The analysis of FN1 suggested that this gene is less abundant than predicted by SAGE. FN1 is a high abundance gene with many reliable tags suggesting transcript variants for FN1. Indeed at least 20 transcripts for FN1 have been described and up to 50 have been predicted (NCBI) (Alberts et al., '94) and the distribution of all the FN1 tags in other SAGE libraries appears complex (SAGEmap). There are several possible explanations for the apparent discrepancy between the observed abundance of fibronectin tags in the SAGE library and the relative abundance as determined by RT-PCR. The primary tag analysis shows that the major FN1 tag is generated from the most 3' N*la* III site of the reference sequence. Minor tags are present but below the level of contamination calculated in CHAPTER 4. However, the amplicon for FN1 could not be designed over the primary tag or near the 3' end of the major transcript due to high levels of repetitive sequences and autologous stretches of single bases. These two conditions and the large variation in characterised transcripts from the FN1 gene, some 20 characterised and some 50 predicted, may conspire to create differences between the SAGE analysis and the PCR analysis.

Two errors arise from the presence of multiple transcripts of the same gene, namely the tag cannot discriminate between the transcripts variants with the same 3' end and the PCR amplicon may span a region in a minor transcript not present in a major transcript. If both these possibilities are true then the tag error will overestimate the level of FN1 relative to the PCR error. Clearly more rigorous characterisation of FN1 transcripts in NHMC is required if these errors are to be fully understood and resolved. Alternatively, the placement of the amplicon could be altered, together with primers that incorporate the tag or tag location. A possible solution for this anomaly would be to redesign the primers over the SAGE tag. This was impossible for FN-1 as it would require changing the PCR cycling parameters, which in turn would interfere with the multiplexing of RT-PCR experiments. This in itself illustrates an issue regarding the analysis of many genes rather than a single gene. As the level of genes available for analysis increases the degree of interrogation of individuals decreases.

The two NHMC transcriptomes were compared and a mathematical estimate of the proportional distribution of 4553 matched pairs was used to assess the similarity between them. The transcriptomes appeared remarkably similar which suggested there was little effect under high glucose compared low glucose. Filtering the raw data was a complex process of calculating probabilities for matched the pair tag pairs, removing redundant tags that matched many genes and resolving tags that matched a few genes and finally selecting candidates to test the reliability of this SAGE analysis to detect differences in transcription. Because of the similarity between the transcriptomes the Beta analysis implied that a large difference in tag frequency or a high degree of sampling was required to confer sufficient probability for further analysis, by convention $p > 0.5$. None of the SAGE tags reached this level of differential abundance or sampling and so the beta parameters were relaxed to favour detection of candidates at the expense of increased false positive data.

When low glucose cultures were compared to high glucose cultures only the two positive controls increased in transcription, TGFβ1 and FN-1, both these genes were present in the SAGE libraries but were excluded from candidature due to low sampling level (TGFβ1) or because there was no significant change. These results suggest, at t = 96h, glucose only increases the transcription of TGFβ1. Failure to detect other genes altered by glucose may be due to inaccuracies in the SAGE analysis process, under-sampling of the libraries or that this is the only effect glucose has on MC. The

transcriptome of the NHMC has not been represented fully and accurately in this SAGE analysis and the effect of glucose on MC is well documented. Given this the differences seen in the SAGE libraries are most likely due to the inherent inaccuracies in sampling and the level of sampling, which concurs with the beta analysis that no significant changes can be detected. The candidates used in these experiments represent the most reliable differential genes in this SAGE analysis but the full list of SAGE candidates has not been tested so there may still be genes that are differentially regulated that have not been tested in these experiments (see APPENDIX 5).

The GeneFilter analyses were able to accurately measure a fraction of the genes that were present on the filter. Again, these experiments suggested that there were very little differences between the high and low glucose cultures. Potential differences were observed in the persistent alteration of 15 genes on the GeneFilter. Of these, eight were selected for further analysis by RT-PCR and two genes concurred with the results from the GeneFilter experiments while one gene displayed up-regulation when down regulation was predicted. APPBP2 showed persistent and significant down regulation in response to glucose and CD58 showed persistent though insignificant up regulation.

The GeneFilter technology proved to be more technically flexible than the SAGE analysis as hybridisation can be used to analyse expanded culture conditions while SAGE is essentially a static profile. The problem of absolute number of targets on the GeneFilter is the most obvious flaw as only 900 genes, one fifth of the potential targets, were accurately quantified.

The NHMC transcriptome contained high levels of the transformation sensitive gene TAGLN. Because of this, a panel of genes was used to investigate the levels of transcription between primary normal MC (NHMC) and their transformed counterparts (HMCL). These experiments revealed striking differences between the two cell types. TAGLN was indeed transformation sensitive with a strong down regulation in HMCL. Interestingly, the largest differences in transcription were observed between collagen isoforms. Types I & III collagen are not normally seen *in vivo* and indeed are very low in HMCL but are highly expressed in NHMC. This was also evident to a lesser extent for type IV collagen, which is normally expressed by MC. This suggests that with regard to these collagens HMCL are more like *in vivo* MC than *in vitro* NHMC. The transformation of MC appeared to have little effect on CTGF and TGBβ1 but large

differences were determined for TGFβ2, SPARC and FN-1. As some of these genes are also important in DN, both *in vivo* and *in vitro* models, recognition of the effects of transformation must be made when substituting HMCL for NHMC *in vitro* or extrapolating results using both NHMC and HMCL into *in vivo* models.

In summary, selections of genes were chosen to evaluate differential transcription based on previous reports, GeneFilter analysis, SAGE analysis and transformation sensitivity. The experiments described quantified the difference between the candidate genes in proliferating cells, quiescent cells under normal glucose, high glucose and osmotic stress. Strong differences were measured between several experimental samples but little change was seen between the experimental samples used to construct the NHMC SAGE libraries.

The most striking results from these experiments revealed a dynamic pattern of differential transcription in response to the removal of serum from the cultures. While this is normal practice in cell culture, used to slow the proliferation of cells, the effect appeared strong for many of the candidates tested. This was not always an up-regulation of transcription, as a suppressive effect was seen for some genes. This phenomenon was not tested across the culture protocol, and many alternative protocols use low levels of serum to 'stabilise' *in vitro* cells. These results suggest further investigation into the response of NHMC to serum would prove useful in determining the extent to which ill-defined serum affects *in vitro* NHMC.

The NHMC was compared to transformed MC (HMCL). It was shown that there were large and important differences between thee two cells. While these data are only qualititative, as the HMCL was not transformed from the NHMC used to compare, there appeared exceptionally large differences between collagen isoforms, TAGLN, SPARC and FN-1. The transformation appeared to have little effect on either the abundance of TGFβ1 TGFβ2 and CTGF, but for TGFβ2 and CTGF appeared to change the level of response to glucose, perhaps increasing the sensitivity to glucose. Further experiments are required to explore this.

All the techniques used to assess the differential transcription of genes in response to high glucose agreed that there was little change. The classic indicator of glucose stress in MC, TGFβ1, was observed to increase in SAGE and RT-PCR. In

SAGE analysis this increase was consistent with predicted magnitude but the sampling level was insufficient to confer likelihood in a beta analysis, and so TGFβ1 was excluded from SAGE candidature. Even selecting candidates calculated to be the most reliable in the SAGE library failed to identify novel genes involved in glucose stress. TGFβ1 was not present on the GeneFilter and so could not be evaluated in those experiments. These results demonstrate that the model is valid and reacting in a predictable way but suggest that the effect of high glucose on transcription in MC is manifest either at a generally low level or outside the time parameters used in this model.

Increasing the level of sampling of the SAGE libraries will have two effects. The transcriptomes of the MC will become more complex as novel tags are sampled, but more importantly, the low level transcription and differences in transcription will become more accurate. All differential transcription identified by GeneFilter analysis was minor and while two differences in expression were observed, only one was statistically significant, yet did not correlate with the SAGE transcriptomes.

CHAPTER 8

8 GENERAL DISCUSSION

8.1 SUMMARY OF RESULTS

The work described in this thesis had two objectives. First, a profile of the transcriptional elements present in NHMC was generated using the method of SAGE. This 'transcriptome' classified all transcriptional elements based on a 10bp 'tag'. It was suggested that the transcriptome of a cell describes its phenotype and this was demonstrated by comparing tags from independent, similarly generated SAGE libraries. The SAGE NHMC transcriptome contained high abundance genes associated with the ECM, cytoskeleton and contractile proteins, which is consistent with the mesenchymal origin and myoblast function of *in vivo* MC. The second objective of this study was to attempt to identify genes altered because of glucose stress. The MC is a target cell in the pathology of diabetes and the *in vitro* culturing of MC in high glucose has become a common model diabetic nephropathy. While SAGE was able to detect many genes implicated in DN, the nature of SAGE meant that it was difficult to accurately identify differentially transcribed genes above the background variations.

8.2 MODEL SYSTEMS

In this particular project the use of a highly defined culture model was preferred for reasons of simplicity and accuracy in the transcriptome analysis. The primary endeavour of this project was to analyse the transcriptome of NHMCs and the model for DN was a useful experiment on which to base such an analysis. The model appeared robust, but failed to demonstrate clear differential transcription in response to glucose. The positive control demonstrated responsiveness to glucose for the cytokine TGFβ1 but all ECM proteins examined changed little in the culture model. This may be due to the short time the cells were exposed to glucose (96hr) or the strict conditions applied to the culture; sub-confluent, serum free and low passage of primary cells. It may equally be likely that an effective *in vitro* model of DN could require contributions from other mechanisms such as mechanical stretch. Given that MC function is associated with the contractile apparatus, it is likely that stretch may mimic hypertension (Harris et al., '92, Gruden et al., '97, Riser et al., '01).

Another possible model for the SAGE analysis in DN could be the micro-dissection of kidneys of rodent models of DM. However, these models would require

exponential increases in the sampling of SAGE libraries with increasing cell complexity in the tissue isolated. While *in vitro* culture models will only ever indicate potential targets for disease, until a clear method for resolving the technical issues surrounding tissue capture, mRNA isolation and library complexity, a pure cell culture remains the preferred model for SAGE analysis.

8.2.1 CULTURE MODEL

Establishing a viable model is important in any global analysis. The degree to which the model can be defined will only increase the quality of the data that is generated from such a model. Central to the aims of this thesis was the ability to detect and quantify genes that were present in cultured glomerular mesangial cells, and candidate genes that changed transcription in response to glucose. The *in vitro* culture of NHMCs was used as a model for DN and it was expected that the SAGE analysis would identify novel targets in DN. It had been reported that irrespective of the proliferative state of MC, they increase production of ECM components and certain factors when cultured under high glucose. This SAGE analysis failed to demonstrate the increased expression of matrix components and was unable to identify novel candidates. Because the positive control for this model, TGFβ1, was shown to alter transcription in response to glucose, it was concluded that the inability of SAGE to identify candidates was due to sampling level, candidate selection or the culture protocol.

8.2.2 CULTURE CONDITIONS

Advances in the *in vitro* culturing of mesangial cells has lead to increased information about the nature of these cells. Several points regarding the culturing of mesangial cells are relevant to an appreciation of experimental procedures. The first point concerns the biological response of mesangial cells in culture, a common issue for all *in vitro* culturing. Removing cells from a biologically defined 3-D matrix and forcing them to proliferate, something for which mesangial cells show a low rate *in vivo*, on a plastic or simple structure is likely to encourage a shift in phenotype. The ECM is no longer regarded as an inert mechanical scaffold but a dynamic biological scaffold, capable of transducing signals, sequestering growth factors and sensing mechanical forces (Sugiyama et al., '98, Elbein and Kaushal, '99, Brady et al., '00,

Zaucke et al., '01). The second point concerns the proliferation required for the generation of cells of sufficient quantity and purity for *in vitro* experiments. By ordinary standards, this would require many generations of cellular replications and contrasts starkly with *in vivo* mesangial cells that show a very low rate of proliferation (Pabst and Sterzel, '83). Indeed excess mesangial cell proliferation is associated with renal disease. Finally, the contact that mesangial cells will have with factors present in the serum used in cell culture will be different from those *in vivo*. Some investigators use relatively high concentrations of non-autologous serum (15-20% FBS/FCS) in their culture medium and mesangial cells normally would be in contact with selective components of plasma (reviewed in (Davies, '94). Taken together these factors, while common to all cell culture, provide a strong case for the cautious interpretation of *in vitro* experiments.

The availability of transformed mesangial cells has created a fourth consideration to the culturing of these cells. Stable SV-40/Ha-ras transformed human and mouse mesangial cells have provided investigators with immortalised MC line that appear to display many of the characteristics of shorter lived primary cells (Sraer et al., '96). A simple set of experiments estimated the difference between these two cell types with regard to abundance of collagen isoforms, cytoskeleton genes and a number of cytokines. With regard to the cytokines CTGF and TGFβ1, the NHMC and HMCL were comparable, with similar abundances for these genes in each. However, with regard to collagen isoforms, particularly the fibrillar type I & III, the HMCL contained several orders of magnitude less mRNA than the primary NHMC. A similar, though less dramatic difference was observed for TAGLN, SPARC and TGFβ2. This clearly demonstrates that the two cell types differ in genes other than simply the transformation elements, and it is likely other genes are affected in a similar or contrary fashion.

A final consideration is the culturing of MC in 2D cultures, rather than 3D environments that mimic the *in vivo* arrangement of MC. Culturing MC in hydrated collagen gels or collagen coated culture dishes produce different proliferative properties, indeed phenotypes, than cells grown on plastic alone. Cells grow faster, but to a lower density, develop longer processes stretching into the gel and generally behave more like their *in vivo* counterparts (Harper et al., '84). The MC also show altered responses to mitogens when culture in such a way (Marx et al., '93).

The literature is varied in its reporting of precise culture conditions and so in our experiments a strict culture model was preferred where FBS and insulin were absent and only glucose could induce transcription. The primary cells were low passage and, while they appeared content through the course of the experiments, the distributor advised against passages past P8 and against culturing in serum free conditions for more than 4 days. This time frame should have been sufficient to induce changes in ECM expansion, and the culture conditions were deliberately so strict as to assign any changes purely to the effect of glucose.

8.2.3 VALIDATING CULTURE PROTOCOL

The culture system was validated to assess the application of the strict culture protocol. As SAGE is such a high resolution, transcriptional analysis, the presence of transformation elements in the transcriptome was unwanted, so although immortalised MCs were available, the use of low passage primary cultures was preferred. When culturing NHMC, mitogenic factors like FBS were also removed to avoid any ill-defined mitogenic factors.

Three techniques were used in the validation of the culture protocol. Northern blots were used to grade the quality of RNA that was isolated from the culture, dot blots were used to investigate the transcription of a series of genes previously determined to be altered in the model and finally a commercial micro-array was used to explore the differential transcription of 5000 targets. The results of these experiments appeared to show that the two culture systems, low glucose and high glucose, were valid and altered gene transcription. Northern blots demonstrated good quality RNA was extracted, but the transcription of two genes hybridised to the northern blots (THBS1 and GLUT1) did not alter. Dot blots appeared to concur with previous reports regarding the transcription of TGFβ1, but did not indicate up-regulation of COLLIVα2 or FN-1. Finally, GeneFilter analysis showed that while the two systems were very similar a number of genes appeared persistently altered across two conditions. Tracked candidates were not altering transcription as predicted, but other genes were, which implied this particular culture system represented a valid model to investigate as uncharacterised genes could be identified. The ability to include genes not associated with proposed mechanisms was facilitated by GeneFilter analysis, as they did not require the selection of candidates. Because SAGE identifies genes based on abundance, rather than

candidature, the NHMC project was considered useful project that would create a substantial volume of data.

8.2.4 PURE CELL CULTURE VERSUS TISSUE

Use of a tissue model in place of a pure cell culture models is feasible and may imply a more accurate, physiological analysis, but requires several considerations. The complexity of the SAGE library must take account of any cell heterogeneity. If there is more than one cell type in the starting material then this will increase the complexity of the library in an exponential manner. Essentially the SAGE libraries will contain two transcriptomes and determining the contribution each transcriptome makes to the tag frequency can create problems in analysis. Much time can be used filtering candidates and care should be exercised that data is not masked by the contributions from two or more transcriptomes. Identifying which cell type has produced which gene can be achieved by using *in situ* hybridisation to confirm location. However, identifying the gene in the mixed SAGE library would be impossible if, say, it were significantly increased in a particular cell type that constitutes only 10% of the cell mass (hence SAGE library) when the gene is not altered in the major cell type that represents 70% of the cell mass. This would not be such a problem if a modest number of genes were analysed in the SAGE library or the tissue sample is considered a micro-organ, but SAGE analysis generates large amounts of data so mixing transcriptomes would complicate any project.

Currently the use of tissue has been most successful in the studies of cancer, where excised tumours were relatively homogenous. However, in some instances individual cellular characteristics are required for physiological accuracy. This is particularly true for 'micro-organs' such as the glomerulus or pancreatic islets, where the overall function of the unit is a combination of strict functioning of each cell type. In the glomerulus, the MC cell is certainly a primary cell, but contributions are also made from podocytes and capillary endothelial cells, not to mention the paracrine effects of neighbouring cells (tubular or ductal epithelial cells for example) or infiltrating immune cells. Each of these glomerular cell types have been implicated in the mechanism of glomerular disease and it would be unwise to ascribe any causative mechanisms to a particular cell type based on bulk glomerular SAGE analysis or claim

a full analysis with bulk tissue without investigating the other members of the glomerulus.

There are other important technical issues associated with the use of whole or micro-dissected tissue models. The isolation of large amounts of mRNA required for SAGE would require large amounts of tissue. Obtaining this tissue in anything but an animal model would be problematic without some sort of control for the effect of long periods without perfusion following excision. The current animal models of DM, either chemically induced (e.g. STZ), or a variety of rodent genetic models, should overcome technical problems, but a model will always be a surrogate marker for disease and thus always require further investigation. One modification of SAGE that may address this, termed micro-SAGE, requires 1000 fold less starting material (Virlon et al., '99). Such a reduction in starting material means that biopsy material may be analysed. Nevertheless, even with these technical modifications to SAGE, complexity of the SAGE library remains an important issue.

8.3 ANALYSING TRANSCRIPTOMES

From the experiments and analyses presented in this thesis, it is apparent that all the techniques used to assess gene expression have their advantages and disadvantages. Hybridisation technology is the most accessible technique to the standard laboratory and with the ease of application comes the freedom to simultaneously monitor many thousands of genes at many points. However, micro-arrays are dependent on the imaging technology available and the errors associated with normalisation and data archiving. The requirement of spatial separation means that currently there is a trade between complexity and ability to characterise specific genes. Furthermore, arrays depend largely on predetermined sequence that require active characterisation. SAGE produced the largest amount of data from the NHMC transcriptome, but the time required to generate the data precluded anything but a snapshot of the transcriptome. The data generated was very simple and as such could be manipulated *in silico* and shared across many experimental platforms. Random sampling of transcripts meant that data was generated based on gene abundance rather than predetermined sequence, thus the true levels of gene transcription could also be fully appreciated. Clearly, the transcriptome is more complex than the genome and global analysis will assist in the understanding of the transcription of genes. Nevertheless, the ultimate usefulness of

that data from projects that produce such volumes will depend on accurate annotation and archiving. SAGE is particularly suited to global transcription analysis and data warehousing so should continue to generate important data regarding the transcriptome.

Global transcription analysis of cell systems affords the opportunity to interrogate models at the highest resolution. Three levels of molecular characterisation are currently defined; the genome, containing the essentially static collection of genes, the transcriptome, representing the transcription of these genes, and the proteome, which describes the translation of genes into functional proteins (Burley et al., '99, Lander, '99, Fields, '01). The resources generated by the HGP has meant that the evolution of analysis techniques, such as high density micro arrays and SAGE, permits simultaneous measurement of transcriptomes which are within the reach of the standard molecular biology laboratory. While proteomics has lagged behind a fuller role for this method of global analysis is also expanding, fuelled in part, by the data generated from the HGP and transcriptional analysis (Neubauer et al., '98, Rubin et al., '00, Gerling et al., '03).

8.3.1 SAGE AS A TOOL TO EXPLORE TRANSCRIPTOMES

The use of SAGE analysis has many advantages over standard techniques. The primary advantage is the non-dependence on sequence and the ease with which data can be stored, retrieved and compared across systems and platforms (Velculescu et al., '00). The SAGE method, while requiring advanced technical skills and intensive in laboratory time should be ultimately more productive than EST sequencing projects. The rapid collection of data is directly proportional to sequencing capacity and tens of thousands of tags can be sampled within a few months. More data can be obtained as sequencing technologies advance.

The SAGE library is a rich source of transcriptional information. The technique only requires the anchoring of all tags to a defined position in the transcript, which is generally within 256bp (4^4) of the 3' polyA tail (4^4 is the mathematical average fragment from a 4-base pair restriction enzyme). From this it can be seen that any variation to the 3' end of a transcript, such as premature termination, indicating inefficient transcription or other transcript variants, will be revealed in the SAGE

library. Assigning these transcript variants can prove problematic but by using several SAGE libraries, generated with different anchoring enzymes, or extending the SAGE tag to 17-20bp, a complete picture of transcripts can be determined. In practice, this level of sampling seems unnecessary as the expansion of the SAGE data warehouse identified transcript variants simply by their presence across libraries and persistent mapping to the same genes. This was demonstrated with the analysis of RTN4 and CTGF (CHAPTER 6.3.2). Data generated from the HGP suggest that alternative transcription is the rule rather than the exception, in some instances accounting for 50% of transcripts from genes (Lander et al., '01, Stamm, '02).

The assignment of mapping data represents the first level of a SAGE library, providing the complete list of transcribed genes, their absolute abundance and 3' transcript variants that predominate. The concept that a sampled population is representative is dependent upon complexity and sampling level, and in reality, one can only infer the likelihood that all members have been detected and quantified. This and other SAGE libraries demonstrated that the representation of genes in a transcriptome is neither simple nor equal. Almost 60% of tags sampled map to only 5% of genes but the majority of unique tags were present only once. A complete list of ultra high abundant genes can probably be determined from a sampling level of a few thousand tags while more moderately transcribed genes may be effectively sampled from tens of thousands of tags. One of the largest SAGE analyses to date (2003) sampled 300,000 tags from four libraries of similar origin. This project was still detecting unique tags at the conclusion of sampling (Zhang et al., '97). Clearly, a method of determining sampling probability is required to assess the completeness of any SAGE library. This has been addressed in several papers and is briefly described in CHAPTER 2.8 (Chen et al., '98a). A SAGE library of 40,000 tags seemed appropriate given the time scale of this project and the analysis of the NHMC transcriptome.

8.3.2 TECHNICAL ERRORS IN THE SAGE PROTOCOL

Using a protocol of many steps will inevitably lead to the increased generation of error. As the technical aspects of SAGE require many manipulations of nucleic acids, some creating potential for nucleic acid instability, then care must be undertaken throughout the protocol and a method of testing this validity must be determined. In this project, the use of sub-libraries created from different mRNA preparations and

generated independently assisted in identifying technical flaws. Variation in any one library could be interpreted as some sort of technical failure and thus variation in the culture and SAGE protocol. From experiments regarding stable accumulation of tags (CHAPTER 4.2.2), the efficiency of tag generation (CHAPTER 4.6.1), the presence of contaminating 5' tags (CHAPTER 4.6.2) and the level of linker contamination (CHAPTER 4.6.3), it was apparent that accumulation across the sub-libraries was by and large linear, which indicates a stable experimental platform for SAGE tag sampling, and thus transcriptome generation.

8.3.3 EXPERIMENTAL ERRORS IN THE SAGE ANALYSIS

Measuring technical errors is more straightforward than estimating experimental errors. As SAGE analysis involved the random sampling of complex populations together with various sequencing and mapping protocols, there is opportunity for several non-empirical errors. These were assigned to 1) sampling errors, 2) sequencing errors, 3) non-random DNA, and 4) non-unique tags. Each will conspire to reduce the efficiency of the SAGE analysis.

8.3.3.1 SAMPLING ERRORS

Sampling errors can be estimated by statistical methods, provided there is an account for the abundance classes in an mRNA population. Uniform functions, where all genes (rather than transcripts) have an equal probability of detection are not useful in estimating SAGE sampling. Simple step functions are more useful, yet due to the abundance classes in transcriptomes they imply that a tag present 500 times has a much lower probability of detection than a tag present 501 times. The method for estimating the sampling efficiency used a 'double exponential' estimation based on previous SAGE library distributions (Stollberg et al., '00). Such calculations are used to estimate sampling levels based on the requirements of the project and estimate any sampling error present. This has been used to examine the properties of the SAGE library and was not used in any error estimation in a particular project nor in this thesis. A modest project of 11,000 tags could be used to characterise a transcriptome of 15,000 genes (p (detection) < 0.98) and a transcriptome of 78,000 genes could be effectively sampled with 64,000 tags (p (detection) <0.92). The real transcriptome is likely to be between

these two examples, possibly at the lower end of complexity. In this project, and many others, there is a trade between increasing the sample level, thus reducing the sampling error, and the timescale for the library sampling. Within the SAGE community, there are libraries as small as 2000 tags and several larger than 500,000 tags. Based on the requirements of this project, a transcriptome of 45,000 tags seemed reasonable for gene detection and determining differential transcription.

8.3.3.2 SEQUENCING ERRORS

Sequencing errors have been measured previously and are dependent on the sequencing platform used in the project. In this project, these errors are between 0.7-1% per base (using ABI Prism sequencing chemistry and a 310 Genetic Analyser). Thus, about 4,000 tags could be expected to be errors. As these errors will occur in an essentially random fashion across the tag, they only affect tags for which there is little data, artificially increasing or decreasing frequency by one. Currently the only way to account for these errors is to only consider tags that are present more than once in the SAGE library. This would exclude a large amount of data, but such data would be only qualitative and useful in describing complexity, rather than any quantifiable significance. Removing all tags present only once revealed the core library (4553 matched pairs), which was used to model the Beta function of proportional distribution. This function was useful in assigning probability to matched pair differences across the two libraries.

8.3.3.3 NON-RANDOM DNA

The non-randomness of DNA will always confound a SAGE analysis, as the ability of a tag to represent a gene is flawed if genes share sequences. The possibility of shared sequences between genes offers the possibility for grouping otherwise unrelated genes into some higher order other than overall gene homology or protein sequence. However, many of the redundant tags were either not very complex, e.g. long strings of adenine residues, or they mapped to repetitive sequences, e.g. Alu elements. Any further analysis would still require the resolution of these tags, and as this phenomenon occurred in some 5,000 tags, was beyond the scope of this project. The technique of extending the tag sequence to the 11^{th} bp proved useful in filtering differential candidates (CHAPTER 7), and was able to resolve about 50% of tags that mapped to less

than 4 genes. Even with this example resolving 2,500 tag redundancies was considered unhelpful. There is currently a need for an analysis method that can account for and resolve these errors, but while the fundamental criterion remains, that a single tag of short sequence represents a single gene, it seems unlikely that this will be resolved. Two experimental variations could minimise these errors. One would require the generation of longer tags and the other would involve the use of two libraries, with different anchoring enzymes. Longer tags would confer capacity for greater complexity and using two libraries of different anchoring enzymes would mean a gene could be identified using two independent tags. Both these modifications would require substantial alterations in SAGE, and the doubling of library sampling. Until there are such libraries available, this is conjecture. There is no reason to suggest that all genes will be resolved using these two techniques. Presently, redundant genes were removed from further analysis.

8.3.3.4 NON-UNIQUE TAGS

A similar situation exists for the non-uniqueness of tags, where a single tag did not represent a gene. This may be due to phenomena such as transcript variants or a technical failure in SAGE. Technical failure was tested by investigating the generation of non-primary tags for a selection of genes from the transcriptome and found to be less than 6%. This error of non-uniqueness of tags is not as problematic as the non-randomness of DNA, as is permits the identification of 3' variation. For this reason this error was permitted.

8.3.4 HYBRIDISATION ANALYSIS OF TRANSCRIPTOMES

The relative simplicity of hybridisation technology has led to its the wide-spread use in many laboratories. The complexity in array technology is increasing and it is possible to simultaneously monitor 20,000 genes in a single experiment (Cheung et al., '99). The GeneFilter used in this project, (GF200), while predicted to monitor some 5,000 genes, was in fact only able to monitor a fifth of these accurately. One main technical problem regarded the resolution of hybridisation signals. Strong signals were clearly resolved but the high abundance genes were generally housekeeping genes that were of little interest in this analysis and in any case did not alter much across the experimental conditions. The detection of rare transcripts, generally more interesting,

required extended exposure or increased concentration of labelled probe in the hybridisation. Both these adjustments result in extensive 'bleeding' of strong signals into weaker ones and obscure a surprisingly large amount of neighbouring gene targets. Technology such as this is based on hybridisation and detection of reporter molecules, whether radioactive isotopes or flurochromes. This indirect measure also introduces errors in detection, normalisation and cross platform comparison. Identical independent experiments would be open to experimental errors that would confound the warehousing of expression data, and in the absence of a single technique or 'industry standard', these problems will persist (Bowtell, '99, Larsson et al., '00, Sherlock, '00). Optimising the technical steps of the protocol for these variables presumably produced the most reliable data, but in the experiments presented here the level of complexity was significantly lower than initially expected, and generally on the borderline of the accepted resolution of the technology.

Two serious issues are evident from micro array technology. One regards the level of sensitivity and the other the complexity of analysis. Sensitivity has been refined to the point where many of the 'bleeding' problems mentioned above are eliminated. This has been facilitated by fluorescent markers and high-resolution laser scanning. Thus, the largest challenge for micro array technology concerns the complexity of the arrays. SAGE has the ability to detect transcript variation, but for micro-arrays to achieve this requires the fragmentation of gene sequences and dispersal of targets. Disregarding the problems associated with hybridisation kinetics, this solution has two consequences. First, the use of many co-ordinates assigned to one gene will reduce the complexity of the array as it relies on physical separation of targets and, second, fragmentation requires active selection and interrogation of characterised gene sequences, which increases the time required to generate the array. Despite these issues, an attempt is being made to discriminate between the 3'UTRs of genes on glass gene-chips with up to 40 co-ordinates containing synthetic oligos. It seems likely that as the technology advances so these issues will be addressed (Lipshutz et al., '99, Tomiuk and Hofmann, '01).

8.3.5 *IN SILICO* ANALYSIS OF GLOBAL EXPRESSION DATA

The generation of such large amounts of data has facilitated the evolution of digital access to biological data, *in silico* analysis (Ermolaeva et al., '98, Boguski, '99). The relatively uncomplicated nature of sequence data (4 bp sequence, fragmented coding structure and triplet codons) means that storing such data is straightforward with modest computing facilities. Following an initial 'wet' period of data accumulation, much of the analysis can take place outside the laboratory. Indeed this has been taken further with the analysis of complex expression profiles purely from the data available in the public domain (McMahon et al., '00, Rana et al., '01, Camargo et al., '01). The exponential growth of genomics data has increased the discovery and annotation of novel genes and the ability to monitor many thousands of gene transcripts simultaneously. All these analyses are possible because of the careful and structured annotation of genomic and transcriptional data and the ease with which it can be accessed.

Prior to SAGE and other technology for global analysis, much of the transcriptional data was collected and measured as transcriptional abundance relative to some sort of constitutively expressed genes such as GAPDH or β-actin. This information was also mostly specific to a particular model system or cell type. With the advent of SAGE, a reproducible technique of transcriptional analysis allowed not only the integration of data from any laboratory, but also warehousing of data in a simple, reliable and accessible fashion. These qualities make SAGE an attractive technology for transcriptome analysis and have resulted in a rapid accumulation of transcriptome data from many disease models and cell systems. Thus, it is now possible to actively mine data directly from a data warehouse, or as presented in this thesis, compare SAGE data from libraries generated from different cells or tissues. Experiments using a digital northern implied that the NHMC was more closely related to fibroblasts, astrocytes, epithelial cells and smooth muscle cells than hepatocytes, cerebellum tissue and cardiac muscle. This similarity concurs with the lineage of NHMC from mesenchymal progenitors, as are fibroblasts, astrocytes and smooth muscle cells but not the other cells types. Recently a small SAGE library was sampled from mesenchymal stem cells and suggested that these progenitors contained characteristic transcripts from many of the potential lineages (Tremain et al., '01). It is likely that SAGE will be further used to describe a transcriptional phenotype.

The generation of SAGE data is, however, a snapshot of a transcriptome. Transcriptomes are dynamic links between the genome and proteome and are subject to the many levels of control. Indeed the untimely or inappropriate transcription of a gene is characteristic of many disease models and actively searching for these candidates offers the possibility of new therapeutic targets. This dynamic nature means that there is also an important time component to gene transcription. This must be accounted for when designing a SAGE project. The sampling of the NHMC SAGE libraries was a large investment in resources, and the possibility that candidates were not detected was dependent upon the culture time point at which the SAGE analysis was performed. Again, this will always occur with global analyses that require large investment. Using hybridisation techniques to study the model allows simultaneous analysis and straightforward introduction of a time dimension to the project. The number of experimental points measured is now dependent on the infrastructure used in the protocol. However, a SAGE analysis provides more than simply an opportunity to build a profile of the transcriptome across an experimental time course. The data from the SAGE libraries contains information about gene transcription and expression at possibly the highest resolution and a method for simple, efficient and reliable annotation.

8.4 NHMC TRANSCRIPTOME

The data from these experiments raise an important question regarding the complexity of the genome. The minimal genome paradigm was formulated in an attempt to simplify the gene annotation of the genome and implies that one gene has one function (Mushegian, '99). However, there are a surprisingly smaller number of genes than expected. This evolved into the theory of gene usage, which suggests that a gene product can have a variety of functions depending on the environment in which it is expressed (Ferea and Brown, '99). Taken together these two theories suggest that while there are a relatively low number of genes in the genome and these genes are essentially static, there is a great diversity of expression of these genes. Further, these gene products may not act in a similar fashion simply because they share sequence homology (Burley et al., '99, Caron et al., '01). Transcript variants may be restricted to cell types or differentiation lineages, and thus protein and specific activity may be restricted to cell type also. This generates two questions regarding the restriction of transcription variation. First, what are the mechanisms for transcript heterogeneity and second, how different are the various transcripts from the genes. This is clearly a complex question

and requires the full picture of gene expression, from the genome, through the transcriptome and ultimately the proteome. SAGE analysis also suggests heterogeneity between transcription products, and instead of gene-based classification, SAGE affords classification based on transcriptional units. SAGE analysis of NHMC suggested that certain transcript variants are restricted to mesenchymal linage or phenotype and that other transcript variants are present in the transcriptome.

8.4.1 MAPPING TAGS TO GENES

It became apparent that mapping tags to genes was not straightforward. The basis of SAGE assumed that the 10bp tag could discriminate between 4^{10} individuals and while the mathematical derivation of this is true, application to a dynamic system such as a transcriptome was not so accurate. It appeared that tags were not unique to genes and that some represented more than one gene. These errors (masking phenomena, discussed in CHAPTER 5) reflect the non-randomness of DNA and the non-uniqueness of tags. These phenomena appeared a common property of SAGE transcription analysis. Also, as the number of data increased, it became more difficult to concentrate on one particular gene and thus there is a decrease in the level of resolution of interrogation. However, this data is not lost, so it is a resource for future studies.

Tag redundancy was apparent in approximately 17% of the tag population and most of these tags mapped to more than five genes. Resolving these redundancies would require large investment in time and resources beyond the scope of this project. For this reason, redundant tags were removed from the SAGE library.

The 2° transcriptome still contained the masking of tags by genes. In this case, the redundancies implied that a single gene could generate more than one tag. While strictly an error in mapping, this data contained information regarding the heterogeneity of transcripts from a gene. Application of this theory to two genes, CTGF and RTN4, supported the usefulness of SAGE in identifying alternative transcripts. This SAGE analysis predicted two major transcripts arising from the RTN4 gene and the annotation of two transcript products from the RTN4 locus with the 3' tails in the predicted position for the generation of the two SAGE tags supports this. CTGF predicted three tags (thus transcripts) from the CTGF gene. Two tags could be rationalised by a

premature transcription termination, but the third appeared to result from anti-sense transcription. The fact that this 'anti-sense' tag is present in other libraries, and continues to map exclusively to CTGF, suggests that there is a functional basis for this transcript. Further study is required to fully understand the true transcription of CTGF and any functional significance of these putative transcript variants.

In order to create the most accurate transcriptomes, removing redundant tag and gene mapping generated the final non-redundant transcriptome. This transcriptome contained 8,774 entries, somewhat reduced from the 1° transcriptome (20,382) and 2° transcriptome (11,853). The 'nr' transcriptome represents the stringent application of SAGE and thus the most accurate map.

Each transcriptome contributes to the transcriptional analysis of the NHMC and all constitute a resource for future analysis. The 1° transcriptome provides complete reliable mapping but contains redundancies in tags and genes. Redundancies in tags many indicate shared regions of DNA and perhaps offer a mechanism for grouping otherwise unrelated genes. The 2° transcriptome does not contain redundant tags but permits more than one tag to represent the same gene. This gene redundancy indicates alternative transcription products from the same gene. The 'nr' transcriptome contains entries to which the most stringent SAGE criteria are applied, one tag maps to one gene.

Approximately 10% (4,116) of tags failed to match any entry in the reliable database. These 'no-match' tags could be generated, or result from technical error (e.g. sequencing), or they could represent 'new' transcripts. Errors generated from sequencing, resulting in the sampling of 'false tags', can only be estimated and accounted for by removal of single tags from the project. Inefficiency of the generation protocol, resulting in tags that are not generated from the most 3' AE site were measured to be below 6%. It may be possible that a percentage of 'no-match' tags could represent completely uncharacterised genes or gene products. Thus, tags that are present more than 3 times and less than 6% of total mapping to one gene, are likely to be real and represent uncharacterised or unreliably mapped gene products. As the HGP progresses, uncharacterised genes and gene products will become less frequent. Throughout the project, the entire set of 'no-match' tags were periodically re-queried in the database. Steadily, mapping data emerged for these tags, reflecting the continued expansion of tag mapping and UniGene clustering. Even so, a current analysis of over

100 SAGE libraries found 375,000 unique tags from some 4.5 million (Chen et al., '02). Of these, 141,000 mapped to UniGene clusters but 234,000 did not. Measuring the potential errors in SAGE and cloning a sample of the no match tags revealed about 70% of them were novel transcripts. These authors suggested that these tags might be a mix of novel genes, novel ORFs or some regulatory function such as RNA interference (RNAi).

Even with the current SAGE mapping data, there are still tags that do not match UniGene clusters. Recent projects all have a level of no-match tags, sometimes 20%(El-Meanawy et al., '00, Yano et al., '00, Velculescu et al., '00). Attempts are being made to address this issue and, rather than the construction of a cluster map, as with UniGene, detailing transcriptional products from a gene that would be more useful for a SAGE project. The generation of a more accurate mapping database is a serious issue for SAGE. Current estimations regarding the contamination of EST libraries by genomic DNA suggest that it is more widespread than predicted and affects about 40% of clusters (Sorek and Safer, '03). This will create a high degree of error in both UniGene clustering and tag mapping, clustering rare transcripts into false genes and poisoning real gene clusters with erroneous ESTs. Investigating methods of minimising these particular errors are underway, but are still in their infancy (reviewed in (Sherlock, '00)).

8.4.2 VALIDATION OF TRANSCRIPTION

The NHMC transcriptome was validated in three ways. A qualitative and semi-quantitative approach, using simple dot blot hybridisations, confirmed the presence of selected genes. Normalisation to GAPDH assessed the relative abundances and only one gene of 29 failed to generate a hybridisation signal above the background. Generally, all the other genes displayed an increasing hybridisation signal in proportion to increasing SAGE determined abundance ($r = 0.84$). Several outliers were present and could represent two further errors. First is the overestimation of the hybridisation signal by cross reactivity or non optimal kinetics, and second is the under-estimation of tag abundance caused by alternative transcript products or redundancies in mapping.

Real time RT-PCR determined a quantitative evaluation of SAGE transcriptome analysis. The correlation of rtRT-PCR data to SAGE tag abundance was also high ($r =$

0.85). This implied that the rtRT-PCR analysis reflected the hybridisation experiments. Possible mechanisms that would cause an overestimation on PCR amplification may relate to transcript heterogeneity and the placement of primers. Under representation by PCR may indicate the incorrect mapping of tags. As SAGEmap is dependent on the dynamic UniGene clustering program, mapping information is evolving with UniGene. Currently, 'builds' for UniGene occur every month and through the course of the project it became apparent that the mapping of some tags was also changing. Clearly, the mapping databases constructed locally require periodical updating for the most reliable mapping. Also, excluded redundant tags that truly map to candidate genes, but were removed due to inability to accurately resolve, may also contribute to an underestimation of true abundance. Resolving the level of contribution of these redundant tags to the specific abundance would require the contribution of each gene to the tag level, which is technically complex.

8.4.3 THE NHMC TRANSCRIPTOME

Once the tags mapped to genes, the transcriptome could be broken down for functional classification, or gene ontology. The groups included in this thesis were structural association, ECM, transcription and translation, enzymes, receptors and markers, cytokines and growth factors, proteases and chaperone proteins. There was a degree of subjective classification as some genes could belong to more than one group. Many of the candidate genes, tracked because of their reported involvement in DN, were not present in these groups, which were essentially composed of high abundance genes. For continuity, these genes grouped together under a section of functional significance for this project (TABLE 6.9).

Genes that were present in the functional groups were all of very high abundance and were present in many other SAGE libraries. This is not surprising, as the high abundance genes would contain the so-called house keeping genes that are constitutively expressed across many cell types and experimental platforms at high levels. What did emerge from these tables were genes that appeared either restricted to mesenchymal cells, like TAGLN and Calponin, or smooth muscle isoforms of contractile apparatus, such as Myosin lc, tropomyosin and α2 smooth muscle actin. In addition, type IV (or basement membrane specific) collagen was present at high abundance, which belies the functional association of NHMCs with the glomerular

basement membrane. The observations support the hypothesis that transcriptome analysis can describe the phenotype of a cell based on the genes it expresses.

Of particular interest in this study was a subset of genes that have been characterised in *in vitro* models of DN and whose candidature in DN is documented. Almost all the genes described in (CHAPTER 1) were present in this SAGE library but apart from a handful, FN-1, COLLIVα2, CTGF, and IGFBPs, they were all present at such low levels they could only be useful in transcriptome annotation and mapping data. Any difference in tag frequency seen with these genes was insignificant in this sample population. Genes for which there was a large amount of data showed no change in their transcription between the two conditions. RT-PCR experiments investigating their transcription also showed that transcription was unaltered across three independent culture series. Despite this, the mapping of the NHMC SAGE library supported firstly that high abundance genes had a high tag frequency and secondly genes were present that differentiated NHMCs from other cell types.

8.4.4 MAPPING ANOMALIES

To test the sensitivity of SAGE in detecting alternative transcripts from the same gene, the data was mined for relationships between tag groups and genes using two sets of tags. Certainly, alternative transcription and mRNA stability have been described in similar *in vitro* models of DN (Abdel Wahab et al., '98, Wahab et al., '00, Holmes et al., '99). In this way SAGE may offer insights into both differential transcription and transcription regulation.

Mapping data for RTN4 revealed two tags present at a high level. Mining SAGEmap for all tags for RTN4 returned a dataset that is composed primarily of four tags. The 1° tag mapped to a characterised cDNA and the remaining three tags mapped to three EST clusters. Interestingly the 1° tag for RTN4 was not present in any other SAGE library including this NHMC library, which could suggest an error in the extraction or mapping of this tag. The remaining three tags mapped to two dense clusters of ESTs, and a medium density cluster of ESTs. SAGE suggested different transcripts for this gene and the EST cluster data supported this. Examining the locus for RTN4 reveals annotations for two genes, one of which extends 600bp 3' to the termination of the other (FIGURE 6.1). Examining the DNA sequence for this 3' area

revealed an AE site within this 600bp tail. Interestingly the 2° and 3° tags accurately mapped to the RTN4 mRNA and the 600bp 3' tail, while the 1° tag was further 3' again. The 4° tag for RTN4 mapped to a previous exon suggesting still further transcript variation. From this analysis, it appears that SAGE is sensitive to alternative transcripts from the same gene.

Examining the NHMC mapping data for the gene CTGF produced similar results in predicting a premature termination of CTGF transcription (or post transcriptional splicing of the 3' tail), but also predicted the generation of an anti-sense tag at significantly high levels. The tag appeared real as it was of high abundance and present in many other SAGE libraries at similar levels (see FIGURE 6.2). Peculiar *cis*-acting elements have been described in the 3'UTR region of CTGF, from where all these tags originate. Perhaps the generation of these tags in some way reflects these elements or they represent an RNAi pathway. Clearly further investigation of CTGF transcription and control of transcription are required before a complete understanding becomes apparent. SAGE appeared to be able to reveal alternative and novel transcription kinetics.

8.4.5 VIRTUAL NORTHERN

The digital nature of SAGE data permits easy comparison across experimental models and platforms. Comparing independent SAGE libraries assessed the quality of the NHMC library. The presence of a particular tag in other independent SAGE libraries supported the validity of this SAGE library. Once validated, other questions regarding expression profiles could further characterise SAGE libraries. It has long been known that relative transcription of many high abundance genes vary little (e.g. housekeeping genes). In addition, cell phenotype is understood to result from the ordered and timely expression of differentiation specific genes. With this in mind three queries were proposed and a virtual Northern constructed from a selection of independent SAGE libraries.

The first question concerned very high abundance genes. In the digital Northern, nearly all these high abundance genes were associated with the cytoskeleton, translation and transcription factors (house-keeping genes). All these genes appeared at

similar levels in all the libraries, supporting the notion that transcription varies little across cell types and these high abundance genes are generally housekeeping genes.

The second question concerned the restriction of genes to cell lineage, in this case mesenchymal. It was predicted that lineage specific genes would be present in this digital northern and there appeared a set of genes that was restricted to mesenchymal cells, particularly TAGLN, myosin lc2, Calponin, leimodon1 and type IV collagen.

Finally, the virtual northern was queried for genes expressed exclusively in NHMC. This cell-restricted digital northern returned only a handful of genes. Almost half were uncharacterised EST clusters, which may indicate uncharacterised genes or gene products specific to NHMC. The remaining genes, particularly activin A and laminin S, may represent gene products or splice variants that are specific to the state of differentiation. While differences were detected, clarification is required.

8.5 DIFFERENTIAL TRANSCRIPTION

The evidence in the literature clearly indicates that high glucose is sufficient to alter the transcription of genes and so this is likely to be the case. The inability of this SAGE analysis to detect changes in the classic ECM genes, generally of high abundance and present in these libraries, may indicate that this culture protocol is not an effective model for DN.

Taking the GeneFilter and SAGE analysis results together provides two sets of conclusions. First regards the similarity between the two experimental systems. Almost all tracked candidates show there was no altered transcription in this system; the exception was TGFβ1, used as a positive control. Second is the implication on the level of sampling in this SAGE project. This SAGE project demonstrated that all the high abundance genes remained at essentially the same level in the experimental sets. The medium to low abundance genes appeared more variable in abundance but this variation was statistically insignificant. In the absence of very high changes in gene transcription, more sampling is required to infer more subtle changes (as with RPS6) or detect changes in low abundance genes. Clearly changes in transcription were occurring, particularly with TGFβ1, but the sampling level requires expanding some 5-fold for this

to be considered significant. This increase in sampling represents a significant expansion of project resources and would more than likely require a more efficient technique. Optimising the molecular cloning and sequencing of tags can most easily achieve increased efficiency. Cloning of tags could conceivably achieve levels of 30-35 tags per clone, and provided the linker and ditag contamination could also be minimised, sequencing of 3000 clones could produce a transcriptome of 100,000 tags. Current high throughput sequencing could resolve this level of sampling in months.

Identifying genes differentially transcribed in response to stimuli identifies genes for further study or possible therapeutic intervention. In this project, this involved subjecting two cultures of NHMCs to low insulin and, physiological or hyperglycaemic concentrations of glucose. Inferred to be an *in vitro* model for diabetic nephropathy, the hypothesis was proposed that high glucose was sufficient for NHMCs to increase the deposition of ECM components and secrete pro-fibrotic cytokines like TGFβ1. Because this is a global analysis it was suggested that transcription of novel genes will also be detected and thus contribute to the understanding of DN.

Identifying genes that responded solely to high concentrations of glucose was the endpoint of this project. Two techniques were used to identify differentially regulated genes. Both proved useful in different ways, and identified potentially differentially transcribed genes, however none of the selected genes altered as assessed using RT-PCR. Therefore, based on the analysis in this model, high concentrations of glucose had no detectible affect on transcription in NHMC, thus the hypothesis must be rejected. However, variation of transcription was detected (e.g. TGFβ1) and the evidence from the literature overwhelmingly demonstrates that glucose is sufficient to alter transcription and so this is likely to remain the case. Possible reasons for inaccuracies will be addressed below.

8.5.1 GeneFilter Analysis

Micro array technology was found to be the most straightforward technique able to rapidly identify a subset of genes that appeared altered in each experimental group. However, the data from these experiments was confined to a random selection of 5000 targets that were immobilised to a nylon filter. Of these, only some 20% provided signals considered strong enough for efficient quantitation and of these, only 10

candidate genes were identified. The persistently altered genes were predominantly uncharacterised ESTs. Thus, while micro-array analysis was a straightforward technique with well-established methodology, it appeared to be restricted in the complexity for gene targets available for analysis. In conclusion, the GeneFilter analysis suggested there was little difference between the two experimental groups, and thus glucose had little effect of the transcription in NHMC.

8.5.2 SAGE ANALYSIS

SAGE has been used to identify differentially transcribed genes in a number of systems and is currently generating a large amount of gene annotation data. The advantage of SAGE over hybridisation technology is, in addition to efficient quantitation of gene abundance, tracking a particular gene in a series of libraries is essentially a digital process. Such tracking can be achieved rapidly and with minimal computing power. This has led to SAGE being used to create a large repository of gene expression data in an extremely short space of time. Using SAGE to identify novel genes in experimental systems provides two levels of data. First, the identification of genes based on their actual presence rather than a rationalised or presumed presence and secondly, a high-resolution transcription profile of an experimental system is generated, the latter containing information on the differential transcription of genes.

In this project, two SAGE libraries were constructed from the *in vitro* model of DN. Matched tag pairs were directly compared and they appeared remarkably similar, which concurred with the GeneFilter analysis. The level of similarity had two ramifications for differential identification. First, the filtering system used to separate significantly differentiated genes from insignificant variations suggested that a large degree of difference was required for significance. Secondly, tracking previously identified candidate genes suggested that they were not altered in this model system. In order to favour discovery over accuracy the parameters for the significance were relaxed and a list of candidates was constructed. This list was presumed to contain an undetermined level of false positives, but as the rtRT-PCR technique was technically straightforward, these could be tested rapidly. The set of candidate genes actively tracked in the SAGE libraries gave two conclusions. First, the high abundance genes demonstrated little variation, and the genes present at low level, generally more interesting, were insufficiently sampled so any observed variation was insignificant.

Of the 14,953 tags detected, some 273 displayed an altered transcription of ±3 fold. Of these, 134 were repressed in high glucose and 139 were increased. 197 tags mapped to characterised mRNA while 60 tags mapped to EST clusters and 16 failed to map. Assessing gene redundancy filtered the candidates and tags from the 1° transcriptome that showed a high level of altered abundance, but were themselves redundant for genes. None of theses candidates were resolved by mapping the 11^{th} bp of the tag and so removed from analysis. Once filtered, a random selection of genes were analysed by rt RT-PCR. Only the positive control, TGFβ1, showed persistent differential transcription between low and high glucose across three independent culture series. This reflected the mathematical inferred likelihood that none of the selected genes was differentially transcribed in this system.

The level of sampling could also account for the exclusion of other candidate genes. Reduced expression of ribosomal protein S6 has been implicated in cell cycle arrest at the G1/S transition. (Chen and Ioannou, '99, Frodin and Gammeltoft, '99). RPS6 is present in these SAGE libraries and is reduced 1.7 fold in the HG libraries (LG n = 32, HG n = 19), which could imply a down regulation consistent with cell cycle arrest. The sampling level for this gene was high, but the level of difference for this gene was low (below ±3) and so it was excluded from further analysis. As there is a large amount of data for this gene (51 tags) there is a greater likelihood of the observed repression being real. This may represent a set of candidates that were overlooked due to more subtle changes in transcription

Along with SAGE determined candidates, a panel of actively tracked genes was compiled based on reported response to glucose. Of these, only the classic indicator of DN was altered in this model system. SAGE predicted a three-fold increase in TGFβ1 expression and RT-PCR determined a 1.8 fold increase. Interestingly the companion markers, collagen, CTGF and other matrix factors showed no significant change in transcription. This suggests that TGFβ1 induction occurs prior to other gene transcription. The real-time RT-PCR analysis supports the SAGE data and the calculated probability of differential tag abundance that the two experimental systems are remarkably similar. This does not mean that transcription is not altered; simply that SAGE was unable to identify changes based on this sample. The hybridisation

experiments in the preliminary experiments and micro array experiments also support this observation.

8.6 CONCLUSIONS OF PROJECT

A transcriptome of cultured NHMC was created using SAGE. This transcriptome sampled 43,000 tags (20,000 unique tags), and reliably mapped them to 11,853 UniGene clusters. As with other SAGE libraries the majority of tags represented about 9% of the genes, while the bulk of the tag mass (70%) was present at a low level (1 tag).

Comparing the transcriptome to other SAGE libraries demonstrated that the NHMC was more similar to cells of mesenchymal origin such as fibroblasts, astrocytes and smooth muscle cells than they were to hepatocytes, cardiac muscle and epithelial cells. These similarities were calculated using transcriptomes of 28,000 unique tags and so are likely to reflect an accurate correlation.

Despite sampling the SAGE libraries to 43,000 tags, approximately one half from cells grown under physiological concentrations of glucose and one half cultured under high glucose, it was difficult to reliably identify novel genes regulated by glucose. Inability to identify differentially regulated genes was probably due to the level of sampling or candidate selection as the evidence in the literature overwhelmingly supports the theory that glucose is sufficient to cause differential gene transcription in cultured MC.

8.7 THESIS

Based of the data generated from the NHMC transcriptome, this SAGE analysis identified a subset of genes that was indicative of the contractile nature of the NHMC and correlation analysis with other transcriptomes grouped cells for the same lineage. Thus, SAGE analysis can be used to describe and classify cells based purely on their transcriptional profile.

SAGE analysis failed to identify novel genes regulated by glucose in NHMC and so it must be concluded that based on this analysis glucose has only a small affect on cultured NHMC. However, a subset of probable candidate genes were generated which require further analysis. This set of data that may still contain novel candidate genes involved in DN.

Bibliography

REFERENCES

Abdel Wahab, N., Gibbs, J. and Mason, R. M. (1998). Regulation of gene expression by alternative polyadenylation and mRNA instability in hyperglycaemic mesangial cells, *Biochem J,* 336 (Pt 2), 405-11.

Adams, M. D. (1996). Serial analysis of gene expression: ESTs get smaller, *Bioessays,* 18, 261-2.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1994) *Molecular Biology of the Cell,* Garland Publishing, New York.

Amiri, F. and Garcia, R. (1999). Regulation of angiotensin II receptors and PKC isoforms by glucose in rat mesangial cells, *Am J Physiol,* 276, F691-9.

Andersen, A. R., Christiansen, J. S., Andersen, J. K., Kreiner, S. and Deckert, T. (1983). Diabetic nephropathy in Type 1 (insulin-dependent) diabetes: an epidemiological study, *Diabetologia,* 25, 496-501.

Anderson, S. S., Kim, Y. and Tsilibary, E. C. (1994). Effects of matrix glycation on mesangial cell adhesion, spreading and proliferation, *Kidney Int,* 46, 1359-67.

Ayo, S. H., Radnik, R., Garoni, J. A., Troyer, D. A. and Kreisberg, J. I. (1991a). High glucose increases diacylglycerol mass and activates protein kinase C in mesangial cell cultures, *Am J Physiol,* 261, F571-7.

Ayo, S. H., Radnik, R. A., Garoni, J. A., Glass, W. F. d. and Kreisberg, J. I. (1990). High glucose causes an increase in extracellular matrix proteins in cultured mesangial cells published erratum appears in Am J Pathol 1990 Aug;137(2):preceding 225, *Am J Pathol,* 136, 1339-48.

Ayo, S. H., Radnik, R. A., Glass, W. F. d., Garoni, J. A., Rampt, E. R., Appling, D. R. and Kreisberg, J. I. (1991b). Increased extracellular matrix synthesis and mRNA in mesangial cells grown in high-glucose medium, *Am J Physiol,* 260, F185-91.

Ayoubi, T. A. and Van De Ven, W. J. (1996). Regulation of gene expression by alternative promoters, *Faseb J,* 10, 453-60.

Barbosa, J. A., Mentzer, S. J., Kamarck, M. E., Hart, J., Biro, P. A., Strominger, J. L. and Burakoff, S. J. (1986). Gene mapping and somatic cell hybrid analysis of the role of human lymphocyte function-associated antigen-3 (LFA-3) in CTL-target cell interactions, *J Immunol,* 136, 3085-91.

Barnes, D. J., Pinto, J. R., Viberti, G., Cameron, J. S., Grunfeld, J. P., Kerr, D. N. S., Ritz, E. and Winearls, C. G. (1998) *The Patient with Diabetes Mellitus,* Oxford University Press, Oxford.

Bassuk, J. A., Pichler, R., Rothmier, J. D., Pippen, J., Gordon, K., Meek, R. L., Bradshaw, A. D., Lombardi, D., Strandjord, T. P., Reed, M., *et al.* (2000). Induction of TGF-beta1 by the matricellular protein SPARC in a rat model of glomerulonephritis, *Kidney Int,* 57, 117-28.

Baynes, J. W. (1991). Role of oxidative stress in development of complications in diabetes, *Diabetes,* 40, 405-12.

Baynes, J. W. and Thorpe, S. R. (1999). Role of oxidative stress in diabetic complications: a new perspective on an old paradigm, *Diabetes,* 48, 1-9.

Bernard, K., Auphan, N., Granjeaud, S., Victorero, G., Schmitt-Verhulst, A. M., Jordan, B. R. and Nguyen, C. (1996). Multiplex messenger assay: simultaneous, quantitative measurement of expression of many genes in the context of T cell activation, *Nucleic Acids Res,* 24, 1435-42.

Bishop, J. O., Morton, J. G., Rosbash, M. and Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA, *Nature,* 250, 199-204.

Black, J. D. (2000). Protein kinase C-mediated regulation of the cell cycle, *Front Biosci,* 5, D406-23.

Blackstock, W. P. and Weir, M. P. (1999). Proteomics: quantitative and physical mapping of cellular proteins, *Trends Biotechnol,* 17, 121-7.

Boguski, M. S. (1999). Biosequence exegesis, *Science,* 286, 453-5.

Bowtell, D. D. (1999). Options available--from start to finish--for obtaining expression data by microarray published erratum appears in Nat Genet 1999 Feb;21(2):241, *Nat Genet,* 21, 25-32.

Bradshaw, A. D., Francki, A., Motamed, K., Howe, C. and Sage, E. H. (1999). Primary mesenchymal cells isolated from SPARC-null mice exhibit altered morphology and rates of proliferation, *Mol Biol Cell,* 10, 1569-79.

Brady, H. R., McGinty, A. and Adler, S. (2000) In *Brenner & Rectors The Kidney*, Vol. 1 (Ed, Brenner, B. M.) W. B. Saunders, Philadelphia, pp. 192-214.

Bron, A. J., Sparrow, J., Brown, N. A., Harding, J. J. and Blakytny, R. (1993). The lens in diabetes, *Eye,* 7 (Pt 2), 260-75.

Brown, S., McGrath, M. J., Ooms, L. M., Gurung, R., Maimone, M. M. and Mitchell, C. A. (1999). Characterization of two isoforms of the skeletal muscle LIM protein 1, SLIM1. Localization of SLIM1 at focal adhesions and the isoform slimmer in the nucleus of myoblasts and cytoplasm of myotubes suggests distinct roles in the

cytoskeleton and in nuclear-cytoplasmic communication, *J Biol Chem,* 274, 27083-91.

Brownlee, M. (1992). Glycation products and the pathogenesis of diabetic complications, *Diabetes Care,* 15, 1835-43.

Brownlee, M. (1995). Advanced protein glycosylation in diabetes and aging, *Annu Rev Med,* 46, 223-34.

Brownlee, M. (2001). Biochemistry and molecular cell biology of diabetic complications, *Nature,* 414, 813-20.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. and Swaminathan, S. (1999). Structural genomics: beyond the human genome project, *Nat Genet,* 23, 151-7.

Camargo, A. A., Samaia, H. P., Dias-Neto, E., Simao, D. F., Migotto, I. A., Briones, M. R., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., *et al.* (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome, *Proc Natl Acad Sci U S A,* 98, 12103-8.

Cameron, N. E., Eaton, S. E., Cotter, M. A. and Tesfaye, S. (2001). Vascular factors and metabolic interactions in the pathogenesis of diabetic neuropathy, *Diabetologia,* 44, 1973-88.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., *et al.* (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains, *Science,* 291, 1289-92.

Cha, D. R., Kim, N. H., Yoon, J. W., Jo, S. K., Cho, W. Y., Kim, H. K. and Won, N. H. (2000). Role of vascular endothelial growth factor in diabetic nephropathy, *Kidney Int Suppl,* 77, S104-12.

Chase, H. P., Jackson, W. E., Hoops, S. L., Cockerham, R. S., Archer, P. G. and O'Brien, D. (1989). Glucose control and the renal and retinal complications of insulin-dependent diabetes, *Jama,* 261, 1155-60.

Chen, F. W. and Ioannou, Y. A. (1999). Ribosomal proteins in cell proliferation and apoptosis, *Int Rev Immunol,* 18, 429-48.

Chen, H., Centola, M., Altschul, S. F. and Metzger, H. (1998a). Characterization of gene expression in resting and activated mast cells published erratum appears in J Exp Med 1998 Dec 21;188(12):2387, *J Exp Med,* 188, 1657-68.

Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J. D. and Wang, S. M. (2002). Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags, *Proc Natl Acad Sci U S A,* 99, 12257-62.

Chen, K. H., Harris, D. L. and Joyce, N. C. (1999). TGF-beta2 in aqueous humor suppresses S-phase entry in cultured corneal endothelial cells, *Invest Ophthalmol Vis Sci,* 40, 2513-9.

Chen, M. S., Huber, A. B., van der Haar, M. E., Frank, M., Schnell, L., Spillmann, A. A., Christ, F. and Schwab, M. E. (2000). Nogo-A is a myelin-associated neurite outgrowth inhibitor and an antigen for monoclonal antibody IN-1, *Nature,* 403, 434-9.

Chen, S. H., Zhou, S., Tan, J. and Schachter, H. (1998b). Transcriptional regulation of the human UDP-GlcNAc:alpha-6-D-mannoside beta-1-2-N-acetylglucosaminyltransferase II gene (MGAT2) which controls complex N-glycan synthesis, *Glycoconj J,* 15, 301-8.

Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R. and Childs, G. (1999). Making and reading microarrays, *Nat Genet,* 21, 15-9.

Chin, E., Zamah, A. M., Landau, D., Gronbcek, H., Flyvbjerg, A., LeRoith, D. and Bondy, C. A. (1997). Changes in facilitative glucose transporter messenger ribonucleic acid levels in the diabetic rat kidney, *Endocrinology,* 138, 1267-75.

Clark, E. A., Golub, T. R., Lander, E. S. and Hynes, R. O. (2000). Genomic analysis of metastasis reveals an essential role for RhoC, *Nature,* 406, 532-5.

Clauss, I. M., Wathelet, M. G., Szpirer, J., Islam, M. Q., Levan, G., Szpirer, C. and Huez, G. A. (1991). Human thymosin-beta 4/6-26 gene is part of a multigene family composed of seven members located on seven different chromosomes, *Genomics,* 9, 174-80.

Claverie, J. M. (2001). Gene number. What if there are only 30,000 human genes?, *Science,* 291, 1255-7.

Cooper, M. E. (1998). Pathogenesis, prevention, and treatment of diabetic nephropathy, *Lancet,* 352, 213-9.

Cortes, P., Zhao, X., Riser, B. L. and Narins, R. G. (1997). Role of glomerular mechanical strain in the pathogenesis of diabetic nephropathy, *Kidney Int,* 51, 57-68.

Couser, W. G. (1993). Pathogenesis of glomerulonephritis, *Kidney Int Suppl,* 42, S19-26.

Davies, M. (1994). The mesangial cell: a tissue culture view, *Kidney Int,* 45, 320-7.

DCCT (1995). The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial, *Diabetes,* 44, 968-83.

DCCT and Group, R. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group see comments, *N Engl J Med,* 329, 977-86.

Death, A. K., Yue, D. K. and Turtle, J. R. (1999). Competitive RT-PCR for measuring metalloproteinase gene expression in human mesangial cells exposed to a hyperglycemic environment, *Biotechniques,* 27, 512-8, 520.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., *et al.* (1998). A physical map of 30,000 human genes, *Science,* 282, 744-6.

Dempsey, E. C., Newton, A. C., Mochly-Rosen, D., Fields, A. P., Reyland, M. E., Insel, P. A. and Messing, R. O. (2000). Protein kinase C isozymes and the regulation of diverse cell responses, *Am J Physiol Lung Cell Mol Physiol,* 279, L429-38.

Du, X. L., Edelstein, D., Dimmeler, S., Ju, Q., Sui, C. and Brownlee, M. (2001). Hyperglycemia inhibits endothelial nitric oxide synthase activity by posttranslational modification at the Akt site, *J Clin Invest,* 108, 1341-8.

Du, X. L., Edelstein, D., Rossetti, L., Fantus, I. G., Goldberg, H., Ziyadeh, F., Wu, J. and Brownlee, M. (2000). Hyperglycemia-induced mitochondrial superoxide overproduction activates the hexosamine pathway and induces plasminogen activator inhibitor-1 expression by increasing Sp1 glycosylation, *Proc Natl Acad Sci U S A,* 97, 12222-6.

Dubin, R. A., Ally, A. H., Chung, S. and Piatigorsky, J. (1990). Human alpha B-crystallin gene and preferential promoter function in lens, *Genomics,* 7, 594-601.

Duncan, M. R., Frazier, K. S., Abramson, S., Williams, S., Klapper, H., Huang, X. and Grotendorst, G. R. (1999). Connective tissue growth factor mediates

transforming growth factor beta-induced collagen synthesis: down-regulation by cAMP, *Faseb J,* 13, 1774-86.

El-Meanawy, M. A., Schelling, J. R., Pozuelo, F., Churpek, M. M., Ficker, E. K., Iyengar, S. and Sedor, J. R. (2000). Use of serial analysis of gene expression to generate kidney expression libraries, *Am J Physiol Renal Physiol,* 279, F383-92.

Elbein, A. and Kaushal, G. P. (1999) In *Medical Biochemistry*(Eds, Baynes, J. and Dominiczak, M.) Mosby, London, pp. Ch26 333-339.

Engerman, R. L., Kern, T. S. and Larson, M. E. (1994). Nerve conduction and aldose reductase inhibition during 5 years of diabetes or galactosaemia in dogs, *Diabetologia,* 37, 141-4.

Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M. and Boguski, M. S. (1998). Data management and analysis for gene expression arrays, *Nat Genet,* 20, 19-23.

Feldman, E. L., Stevens, M. J. and Greene, D. A. (1997). Pathogenesis of diabetic neuropathy, *Clin Neurosci,* 4, 365-70.

Ferea, T. L. and Brown, P. O. (1999). Observing the living genome, *Curr Opin Genet Dev,* 9, 715-22.

Fernandez-Pol, J. A., Klos, D. J. and Hamilton, P. D. (1993). A growth factor-inducible gene encodes a novel nuclear protein with zinc finger structure, *J Biol Chem,* 268, 21198-204.

Fields, C., Adams, M. D., White, O. and Venter, J. C. (1994). How many genes in the human genome? news see comments, *Nat Genet,* 7, 345-6.

Fields, S. (1997). The future is function news, *Nat Genet,* 15, 325-7.

Fields, S. (2001). Proteomics. Proteomics in genomeland, *Science,* 291, 1221-4.

Fioretto, P., Kim, Y. and Mauer, M. (1998a). Diabetic nephropathy as a model of reversibility of established renal lesions, *Curr Opin Nephrol Hypertens,* 7, 489-94.

Fioretto, P., Steffes, M. W., Sutherland, D. E., Goetz, F. C. and Mauer, M. (1998b). Reversal of lesions of diabetic nephropathy after pancreas transplantation see comments, *N Engl J Med,* 339, 69-75.

Fisher, E. J., McLennan, S. V., Yue, D. K. and Turtle, J. R. (1997). High glucose reduces generation of plasmin activity by mesangial cells, *Microvasc Res,* 53, 173-8.

Floege, J., Eng, E., Young, B. A. and Johnson, R. J. (1993). Factors involved in the regulation of mesangial cell proliferation in vitro and in vivo, *Kidney Int Suppl,* 39, S47-54.

Floege, J., Topley, N., Hoppe, J., Barrett, T. B. and Resch, K. (1991). Mitogenic effect of platelet-derived growth factor in human glomerular mesangial cells: modulation and/or suppression by inflammatory cytokines, *Clin Exp Immunol,* 86, 334-41.

Floege, J., Topley, N., Wessel, K., Kaever, V., Radeke, H., Hoppe, J., Kishimoto, T. and Resch, K. (1990). Monokines and platelet-derived growth factor modulate prostanoid production in growth arrested, human mesangial cells, *Kidney Int,* 37, 859-69.

Francki, A., Bradshaw, A. D., Bassuk, J. A., Howe, C. C., Couser, W. G. and Sage, E. H. (1999). SPARC regulates the expression of collagen type I and transforming growth factor-beta1 in mesangial cells, *J Biol Chem,* 274, 32145-52.

Frank, R. N., Hoffman, W. H., Podgor, M. J., Joondeph, H. C., Lewis, R. A., Margherio, R. R., Nachazel, D. P., Jr., Weiss, H., Christopherson, K. W. and Cronin, M. A. (1982). Retinopathy in juvenile-onset type I diabetes of short duration, *Diabetes,* 31, 874-82.

Frodin, M. and Gammeltoft, S. (1999). Role and regulation of 90 kDa ribosomal S6 kinase (RSK) in signal transduction, *Mol Cell Endocrinol,* 151, 65-77.

Frost, M. R. and Guggenheim, J. A. (1999). Mammalian polyadenylation sites: implications for differential display, *Nucleic Acids Res,* 27, 1386-91.

Gerling, I. C., Solomon, S. S. and Bryer-Ash, M. (2003). Genomes, transcriptomes, and proteomes: molecular medicine and its impact on medical practice, *Arch Intern Med,* 163, 190-8.

Gilbert, R. E., McNally, P. G., Cox, A., Dziadek, M., Rumble, J., Cooper, M. E. and Jerums, G. (1995). SPARC gene expression is reduced in early diabetes-related kidney growth, *Kidney Int,* 48, 1216-25.

Giugliano, D., Ceriello, A. and Paolisso, G. (1996). Oxidative stress and diabetic vascular complications, *Diabetes Care,* 19, 257-67.

Glogowski, E. A., Tsiani, E., Zhou, X., Fantus, I. G. and Whiteside, C. (1999). High glucose alters the response of mesangial cell protein kinase C isoforms to endothelin-1, *Kidney Int,* 55, 486-99.

Goldberg, H. J., Scholey, J. and Fantus, I. G. (2000). Glucosamine activates the plasminogen activator inhibitor 1 gene promoter through Sp1 DNA binding sites in glomerular mesangial cells, *Diabetes,* 49, 863-71.

Gosain, A. K., Song, L., Yu, P., Mehrara, B. J., Maeda, C. Y., Gold, L. I. and Longaker, M. T. (2000). Osteogenesis in cranial defects: reassessment of the concept of critical size and the expression of TGF-beta isoforms, *Plast Reconstr Surg,* 106, 360-71; discussion 372.

Greene, D. A., Arezzo, J. C. and Brown, M. B. (1999). Effect of aldose reductase inhibition on nerve conduction and morphometry in diabetic neuropathy. Zenarestat Study Group, *Neurology,* 53, 580-91.

Grotendorst, G. R., Lau, L. F. and Perbal, B. (2000). CCN proteins are distinct from and should not be considered members of the insulin-like growth factor-binding protein superfamily, *Endocrinology,* 141, 2254-6.

Gruden, G., Thomas, S., Burt, D., Lane, S., Chusney, G., Sacks, S. and Viberti, G. (1997). Mechanical stretch induces vascular permeability factor in human mesangial cells: mechanisms of signal transduction, *Proc Natl Acad Sci U S A,* 94, 12112-6.

Ha, H. and Lee, H. B. (2000). Reactive oxygen species as glucose signaling molecules in mesangial cells cultured under high glucose, *Kidney Int,* 58 Suppl 77, S19-25.

Ha, H., Lee, S. H. and Kim, K. H. (1997). Effects of rebamipide in a model of experimental diabetes and on the synthesis of transforming growth factor-beta and fibronectin, and lipid peroxidation induced by high glucose in cultured mesangial cells, *J Pharmacol Exp Ther,* 281, 1457-62.

Hamada, Y., Araki, N., Koh, N., Nakamura, J., Horiuchi, S. and Hotta, N. (1996). Rapid formation of advanced glycation end products by intermediate metabolites of glycolytic pathway and polyol pathway, *Biochem Biophys Res Commun,* 228, 539-43.

Haneda, M., Kikkawa, R., Horide, N., Togawa, M., Koya, D., Kajiwara, N., Ooshima, A. and Shigeta, Y. (1991). Glucose enhances type IV collagen production in cultured rat glomerular mesangial cells, *Diabetologia,* 34, 198-200.

Harper, P. A., Robinson, J. M., Hoover, R. L., Wright, T. C. and Karnovsky, M. J. (1984). Improved methods for culturing rat glomerular cells, *Kidney Int,* 26, 875-80.

Harris, R. C., Akai, Y., Yasuda, T. and Homma, T. (1994). The role of physical forces in alterations of mesangial cell function, *Kidney Int Suppl,* 45, S17-21.

Harris, R. C., Haralson, M. A. and Badr, K. F. (1992). Continuous stretch-relaxation in culture alters rat mesangial cell morphology, growth characteristics, and metabolic activity, *Lab Invest,* 66, 548-54.

Hart, G. W. (1997). Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins, *Annu Rev Biochem,* 66, 315-35.

Hashimoto, S., Suzuki, T., Dong, H. Y., Yamazaki, N. and Matsushima, K. (1999). Serial analysis of gene expression in human monocytes and macrophages, *Blood,* 94, 837-44.

Hasslacher, C., Ritz, E., Terpstra, J., Gallasch, G., Kunowski, G. and Rall, C. (1985). Natural history of nephropathy in type I diabetes. Relationship to metabolic control and blood pressure, *Hypertension,* 7, II74-8.

Hastie, N. D. and Bishop, J. O. (1976). The expression of three abundance classes of messenger RNA in mouse tissues, *Cell,* 9, 761-74.

Hayashi, T., Matsufuji, S. and Hayashi, S. (1997). Characterization of the human antizyme gene, *Gene,* 203, 131-9.

Heilig, C. W., Brosius, F. C., 3rd and Henry, D. N. (1997a). Glucose transporters of the glomerulus and the implications for diabetic nephropathy, *Kidney Int Suppl,* 60, S91-9.

Heilig, C. W., Concepcion, L. A., Riser, B. L., Freytag, S. O., Zhu, M. and Cortes, P. (1995). Overexpression of glucose transporters in rat mesangial cells cultured in a normal glucose milieu mimics the diabetic phenotype, *J Clin Invest,* 96, 1802-14.

Heilig, C. W., Liu, Y., England, R. L., Freytag, S. O., Gilbert, J. D., Heilig, K. O., Zhu, M., Concepcion, L. A. and Brosius, F. C., 3rd (1997b). D-glucose stimulates mesangial cell GLUT1 expression and basal and IGF-I-sensitive glucose uptake in rat mesangial cells: implications for diabetic nephropathy, *Diabetes,* 46, 1030-9.

Heller, R. A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D. E. and Davis, R. W. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays, *Proc Natl Acad Sci U S A,* 94, 2150-5.

Hempel, A., Maasch, C., Heintze, U., Lindschau, C., Dietz, R., Luft, F. C. and Haller, H. (1997). High glucose concentrations increase endothelial cell permeability via activation of protein kinase C alpha, *Circ Res,* 81, 363-71.

Henry, D. N., Busik, J. V., Brosius, F. C., 3rd and Heilig, C. W. (1999). Glucose transporters control gene expression of aldose reductase, PKCalpha, and GLUT1 in mesangial cells in vitro, *Am J Physiol,* 277, F97-104.

Hibi, K., Liu, Q., Beaudry, G. A., Madden, S. L., Westra, W. H., Wehage, S. L., Yang, S. C., Heitmiller, R. F., Bertelsen, A. H., Sidransky, D., *et al.* (1998). Serial analysis of gene expression in non-small cell lung cancer, *Cancer Res,* 58, 5690-4.

Hoffman, B. B., Sharma, K., Zhu, Y. and Ziyadeh, F. N. (1998). Transcriptional activation of transforming growth factor-beta1 in mesangial cell culture by high glucose concentration, *Kidney Int,* 54, 1107-16.

Holmes, D. I., Abdel Wahab, N. and Mason, R. M. (1997). Identification of glucose-regulated genes in human mesangial cells by mRNA differential display, *Biochem Biophys Res Commun,* 238, 179-84.

Holmes, D. I., Wahab, N. A. and Mason, R. M. (1999). Cloning and characterization of ZNF236, a glucose-regulated Kruppel-like zinc-finger gene mapping to human chromosome 18q22-q23, *Genomics,* 60, 105-9.

Horney, M. J., Shirley, D. W., Kurtz, D. T. and Rosenzweig, S. A. (1998). Elevated glucose increases mesangial cell sensitivity to insulin-like growth factor I, *Am J Physiol,* 274, F1045-53.

Hotta, N. (1997). New concepts and insights on pathogenesis and treatment of diabetic complications: polyol pathway and its inhibition, *Nagoya J Med Sci,* 60, 89-100.

Hudson, B. I., Stickland, M. H. and Grant, P. J. (1998). Identification of polymorphisms in the receptor for advanced glycation end products (RAGE) gene: prevalence in type 2 diabetes and ethnic groups, *Diabetes,* 47, 1155-7.

Ihm, C., Park, J., Ahn, J., Lee, T., Cho, B. and Kim, M. (1995). Effect of glucose and cytokines on the expression of cell adhesion molecules on mesangial cells, *Kidney Int Suppl,* 51, S39-42.

Ishii, H., Jirousek, M. R., Koya, D., Takagi, C., Xia, P., Clermont, A., Bursell, S. E., Kern, T. S., Ballas, L. M., Heath, W. F., *et al.* (1996). Amelioration of vascular dysfunctions in diabetic rats by an oral PKC beta inhibitor, *Science,* 272, 728-31.

Ishii, H., Koya, D. and King, G. L. (1998). Protein kinase C activation and its role in the development of vascular complications in diabetes mellitus, *J Mol Med,* 76, 21-31.

Ison, J. C., O'Leary, S. B., Bleasby, A. J. and Moss, D. (2000). The Bioinformatics Resource, *Trends Biochem Sci,* 25, 299.

Isono, M., Mogyorosi, A., Han, D. C., Hoffman, B. B. and Ziyadeh, F. N. (2000). Stimulation of TGF-beta type II receptor by high glucose in mouse mesangial cells and in diabetic kidney, *Am J Physiol Renal Physiol,* 278, F830-F838.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., *et al.* (1999). The transcriptional program in the response of human fibroblasts to serum., *Science,* 283, 83-7.

Jackle-Meyer, I., Szukics, B., Neubauer, K., Metze, V., Petzoldt, R. and Stolte, H. (1995). Extracellular matrix proteins as early markers in diabetic nephropathy, *Eur J Clin Chem Clin Biochem,* 33, 211-9.

James, L. R., Fantus, I. G., Goldberg, H., Ly, H. and Scholey, J. W. (2000). Overexpression of GFAT activates PAI-1 promoter in mesangial cells, *Am J Physiol Renal Physiol,* 279, F718-27.

Kadonaga, J. T., Courey, A. J., Ladika, J. and Tjian, R. (1988). Distinct regions of Sp1 modulate DNA binding and transcriptional activation, *Science,* 242, 1566-70.

Kaneto, H., Xu, G., Song, K. H., Suzuma, K., Bonner-Weir, S., Sharma, A. and Weir, G. C. (2001). Activation of the hexosamine pathway leads to deterioration of pancreatic beta-cell function through the induction of oxidative stress, *J Biol Chem,* 276, 31099-104.

Kendall, M. and Stuart, A. (1977) In *The Advanced Theory of Statistics*, Vol. 1 Macmillan Publishing Co. Inc, New York, pp. Chapter 6.

Kenzelmann, M. and Muhlemann, K. (2000). Transcriptome analysis of fibroblast cells immediate-early after human cytomegalovirus infection, *J Mol Biol,* 304, 741-51.

King, R. H. (2001). The role of glycation in the pathogenesis of diabetic polyneuropathy, *Mol Pathol,* 54, 400-8.

Kitamura, M. and Ishikawa, Y. (1998). Three-dimensional matrix primes mesangial cells to down-regulation of alpha-smooth muscle actin via deactivation of CArG box elements, *Kidney Int,* 53, 690-7.

Klein, R., Klein, B. E., Moss, S. E., Davis, M. D. and DeMets, D. L. (1988). Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy, *JAMA,* 260, 2864-71.

Kolm-Litty, V., Sauer, U., Nerlich, A., Lehmann, R. and Schleicher, E. D. (1998a). High glucose-induced transforming growth factor beta1 production is mediated by the hexosamine pathway in porcine glomerular mesangial cells, *J Clin Invest,* 101, 160-9.

Kolm-Litty, V., Tippmer, S., Haring, H. U. and Schleicher, E. (1998b). Glucosamine induces translocation of protein kinase C isoenzymes in mesangial cells, *Exp Clin Endocrinol Diabetes,* 106, 377-83.

Kothapalli, D. and Grotendorst, G. R. (2000). CTGF modulates cell cycle progression in cAMP-arrested NRK fibroblasts, *J Cell Physiol,* 182, 119-26.

Koya, D., Jirousek, M. R., Lin, Y. W., Ishii, H., Kuboki, K. and King, G. L. (1997). Characterization of protein kinase C beta isoform activation on the gene expression of transforming growth factor-beta, extracellular matrix components, and prostanoids in the glomeruli of diabetic rats, *J Clin Invest,* 100, 115-26.

Kreisberg, J. I., Radnik, R. A., Ayo, S. H., Garoni, J. and Saikumar, P. (1994). High glucose elevates c-fos and c-jun transcripts and proteins in mesangial cell cultures, *Kidney Int,* 46, 105-12.

Kreisberg, J. I., Venkatachalam, M. and Troyer, D. (1985). Contractile properties of cultured glomerular mesangial cells, *Am J Physiol,* 249, F457-63.

Kreppel, L. K., Blomberg, M. A. and Hart, G. W. (1997). Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats, *J Biol Chem,* 272, 9308-15.

Kriz, W. and Kaissling, B. (1992) In *The Kidney: Physiology & Pathophysiology*, Vol. 2 (Eds, Selwin, D. W. and Griebisch, G.) Raven Press Ltd, New York, pp. Ch23.

Krolewski, A. S., Canessa, M., Warram, J. H., Laffel, L. M., Christlieb, A. R., Knowler, W. C. and Rand, L. I. (1988). Predisposition to hypertension and susceptibility to renal disease in insulin-dependent diabetes mellitus, *N Engl J Med,* 318, 140-5.

Krolewski, A. S., Warram, J. H., Christlieb, A. R., Busick, E. J. and Kahn, C. R. (1985). The changing natural history of nephropathy in type I diabetes, *Am J Med,* 78, 785-94.

Kubota, S., Hattori, T., Nakanishi, T. and Takigawa, M. (1999). Involvement of cis-acting repressive element(s) in the 3'-untranslated region of human connective tissue growth factor gene, *FEBS Lett,* 450, 84-8.

Kubota, S., Kondo, S., Eguchi, T., Hattori, T., Nakanishi, T., Pomerantz, R. J. and Takigawa, M. (2000). Identification of an RNA element that confers post-transcriptional repression of connective tissue growth factor/hypertrophic chondrocyte specific 24 (ctgf/hcs24) gene: similarities to retroviral RNA-protein interactions, *Oncogene,* 19, 4773-86.

Kwiatkowski, D. J., Stossel, T. P., Orkin, S. H., Mole, J. E., Colten, H. R. and Yin, H. L. (1986). Plasma and cytoplasmic gelsolins are encoded by a single gene and contain a duplicated actin-binding domain, *Nature,* 323, 455-8.

Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., *et al.* (1999). A public database for gene expression in human cancers, *Cancer Res,* 59, 5403-7.

Lander, E. S. (1996). The new genomics: global views of biology see comments, *Science,* 274, 536-9.

Lander, E. S. (1999). Array of hope, *Nat Genet,* 21, 3-4.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome, *Nature,* 409, 860-921.

Larsson, M., Stahl, S., Uhlen, M. and Wennborg, A. (2000). Expression profile viewer (ExProView): a software tool for transcriptome analysis, *Genomics,* 63, 341-53.

Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J. and Altschul, S. F. (2000). SAGEmap: A public gene expression resource In Process Citation, *Genome Res,* 10, 1051-60.

Lawson, D., Harrison, M. and Shapland, C. (1997). Fibroblast transgelin and smooth muscle SM22alpha are the same protein, the expression of which is down-regulated in many cell lines, *Cell Motil Cytoskeleton,* 38, 250-7.

Ledakis, P., Tanimura, H. and Fojo, T. (1998). Limitations of differential display, *Biochem Biophys Res Commun,* 251, 653-6.

Lee, H. B., Ha, H., Kim, S. I. and Ziyadeh, F. N. (2000). Diabetic kidney disease research: Where do we stand at the turn of the century?, *Kidney Int,* 58, 1-2.

Lehmann, R. and Schleicher, E. D. (2000). Molecular mechanism of diabetic nephropathy, *Clin Chim Acta,* 297, 135-44.

Lennon, G., Auffray, C., Polymeropoulos, M. and Soares, M. B. (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression, *Genomics,* 33, 151-2.

Li, Y. M., Mitsuhashi, T., Wojciechowicz, D., Shimizu, N., Li, J., Stitt, A., He, C., Banerjee, D. and Vlassara, H. (1996). Molecular identity and cellular distribution of advanced glycation endproduct receptors: relationship of p60 to OST-48 and p90 to 80K-H membrane proteins, *Proc Natl Acad Sci U S A,* 93, 11047-52.

Liang, P. and Pardee, A. B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction see comments, *Science,* 257, 967-71.

Lin, B., Rommens, J. M., Graham, R. K., Kalchman, M., MacDonald, H., Nasir, J., Delaney, A., Goldberg, Y. P. and Hayden, M. R. (1993). Differential 3' polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression, *Hum Mol Genet,* 2, 1541-5.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays, *Nat Genet,* 21, 20-4.

Logan, A., Green, J., Hunter, A., Jackson, R. and Berry, M. (1999). Inhibition of glial scarring in the injured rat brain by a recombinant human monoclonal antibody to transforming growth factor-beta2, *Eur J Neurosci,* 11, 2367-74.

Luo, G., Ducy, P., McKee, M. D., Pinero, G. J., Loyer, E., Behringer, R. R. and Karsenty, G. (1997). Spontaneous calcification of arteries and cartilage in mice lacking matrix GLA protein, *Nature,* 386, 78-81.

Madden, S. L., Galella, E. A., Zhu, J., Bertelsen, A. H. and Beaudry, G. A. (1997). SAGE transcript profiles for p53-dependent growth regulation, *Oncogene,* 15, 1079-85.

Makino, H., Kashihara, N., Sugiyama, H., Kanao, K., Sekikawa, T., Shikata, K., Nagai, R. and Ota, Z. (1995). Phenotypic changes of the mesangium in diabetic nephropathy, *J Diabetes Complications,* 9, 282-4.

Makino, H., Yamasaki, Y., Haramoto, T., Shikata, K., Hironaka, K., Ota, Z. and Kanwar, Y. S. (1993). Ultrastructural changes of extracellular matrices in diabetic nephropathy revealed by high resolution scanning and immunoelectron microscopy, *Lab Invest,* 68, 45-55.

Malakoff, D. and Service, R. F. (2001). Genomania meets the bottom line, *Science,* 291, 1193-203.

Marshall, S., Bacote, V. and Traxinger, R. R. (1991a). Complete inhibition of glucose-induced desensitization of the glucose transport system by inhibitors of mRNA synthesis. Evidence for rapid turnover of glutamine:fructose-6-phosphate amidotransferase, *J Biol Chem,* 266, 10155-61.

Marshall, S., Bacote, V. and Traxinger, R. R. (1991b). Discovery of a metabolic pathway mediating glucose-induced desensitization of the glucose transport system. Role of hexosamine biosynthesis in the induction of insulin resistance, *J Biol Chem,* 266, 4706-12.

Marx, M., Daniel, T. O., Kashgarian, M. and Madri, J. A. (1993). Spatial organization of the extracellular matrix modulates the expression of PDGF-receptor subunits in mesangial cells, *Kidney Int,* 43, 1027-41.

Matz, M. V. and Lukyanov, S. A. (1998). Different strategies of differential display: areas of application, *Nucleic Acids Res,* 26, 5537-43.

Matzuk, M. M., Kumar, T. R., Vassalli, A., Bickenbach, J. R., Roop, D. R., Jaenisch, R. and Bradley, A. (1995). Functional analysis of activins during mammalian development, *Nature,* 374, 354-6.

McClain, D. A. and Crook, E. D. (1996). Hexosamines and insulin resistance, *Diabetes,* 45, 1003-9.

McClain, D. A., Paterson, A. J., Roos, M. D., Wei, X. and Kudlow, J. E. (1992). Glucose and glucosamine regulate growth factor gene expression in vascular smooth muscle cells, *Proc Natl Acad Sci U S A,* 89, 8150-4.

McKusick, V. A. (1997). Genomics: structural and functional studies of genomes, *Genomics,* 45, 244-9.

McLennan, S. V., Fisher, E. J., Yue, D. K. and Turtle, J. R. (1994). High glucose concentration causes a decrease in mesangium degradation. A factor in the pathogenesis of diabetic nephropathy, *Diabetes,* 43, 1041-5.

McLennan, S. V., Yue, D. K. and Turtle, J. R. (1998). Effect of glucose on matrix metalloproteinase activity in mesangial cells, *Nephron,* 79, 293-8.

McMahon, R., Murphy, M., Clarkson, M., Taal, M., Mackenzie, H. S., Godson, C., Martin, F. and Brady, H. R. (2000). IHG-2, a mesangial cell gene induced by high glucose, is human gremlin. Regulation By extracellular glucose concentration,

cyclic mechanical strain, and transforming growth factor-beta1, *J Biol Chem,* 275, 9901-4.

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., *et al.* (2001). A physical map of the human genome, *Nature,* 409, 934-41.

Mene, P., Festuccia, F., Polci, R., Pugliese, F. and Cinotti, G. A. (2001). Diabetic nephropathy and advanced glycation end products, *Contrib Nephrol*, 22-32.

Mene, P., Pugliese, F. and Cinotti, G. A. (1996). Regulation of capacitative calcium influx in cultured human mesangial cells: roles of protein kinase C and calmodulin, *J Am Soc Nephrol,* 7, 983-90.

Mene, P., Simonson, M. S. and Dunn, M. J. (1989). Physiology of the mesangial cell, *Physiol Rev,* 69, 1347-424.

Mizisin, A. P., Li, L. and Calcutt, N. A. (1997). Sorbitol accumulation and transmembrane efflux in osmotically stressed JS1 schwannoma cells, *Neurosci Lett,* 229, 53-6.

Mizisin, A. P., Li, L., Perello, M., Freshwater, J. D., Kalichman, M. W., Roux, L. and Calcutt, N. A. (1996). Polyol pathway and osmoregulation in JS1 Schwann cells grown in hyperglycemic and hyperosmotic conditions, *Am J Physiol,* 270, F90-F97.

Moiseeva, E. P., Javed, Q., Spring, E. L. and de Bono, D. P. (2000). Galectin 1 is involved in vascular smooth muscle cell proliferation, *Cardiovasc Res,* 45, 493-502.

Montoliu, L., Rigau, J. and Puigdomenech, P. (1990). Multiple polyadenylation sites are active in the alpha 1-tubulin gene from Zea mays, *FEBS Lett,* 277, 29-32.

Morocutti, A., Sethi, M., Hayward, A., Lee, A. and Viberti, G. (1998). Glutathione reverses the growth abnormalities of skin fibroblasts from insulin-dependent diabetic patients with nephropathy, *J Am Soc Nephrol,* 9, 1060-6.

Mozes, M. M., Hodics, T. and Kopp, J. B. (1999). Isoform specificity of commercially-available anti-TGF-beta antibodies, *J Immunol Methods,* 225, 87-93.

Murphy, G. (1995). Matrix metalloproteinases and their inhibitors, *Acta Orthop Scand Suppl,* 266, 55-60.

Murphy, M., Godson, C., Cannon, S., Kato, S., Mackenzie, H. S., Martin, F. and Brady, H. R. (1999). Suppression subtractive hybridization identifies high glucose

levels as a stimulus for expression of connective tissue growth factor and other genes in human mesangial cells, *J Biol Chem,* 274, 5830-4.

Mushegian, A. (1999). The minimal genome concept, *Curr Opin Genet Dev,* 9, 709-14.

Nagata, M., Scharer, K. and Kriz, W. (1992). Glomerular damage after uninephrectomy in young rats. I. Hypertrophy and distortion of capillary architecture, *Kidney Int,* 42, 136-47.

Nahman, N. S., Jr., Leonhart, K. L., Cosio, F. G. and Hebert, C. L. (1992). Effects of high glucose on cellular proliferation and fibronectin production by cultured human mesangial cells, *Kidney Int,* 41, 396-402.

Nakamura, T., Fukui, M., Ebihara, I., Osada, S., Nagaoka, I., Tomino, Y. and Koide, H. (1993). mRNA expression of growth factors in glomeruli from diabetic rats, *Diabetes,* 42, 450-6.

Nathan, D. M. (1992). The rationale for glucose control in diabetes mellitus, *Endocrinol Metab Clin North Am,* 21, 221-35.

Neeper, M., Schmidt, A. M., Brett, J., Yan, S. D., Wang, F., Pan, Y. C., Elliston, K., Stern, D. and Shaw, A. (1992). Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins, *J Biol Chem,* 267, 14998-5004.

Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A. and Mann, M. (1998). Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex see comments, *Nat Genet,* 20, 46-50.

Nilsson, J., Koskiniemi, S., Persson, K., Grahn, B. and Holm, I. (1997). Polyamines regulate both transcription and translation of the gene encoding ornithine decarboxylase antizyme in mouse, *Eur J Biochem,* 250, 223-31.

Nishikawa, T., Edelstein, D., Du, X. L., Yamagishi, S., Matsumura, T., Kaneda, Y., Yorek, M. A., Beebe, D., Oates, P. J., Hammes, H. P., *et al.* (2000). Normalizing mitochondrial superoxide production blocks three pathways of hyperglycaemic damage, *Nature,* 404, 787-90.

Nugent, P., Ma, L. and Greene, R. M. (1998). Differential expression and biological activity of retinoic acid-induced TGFbeta isoforms in embryonic palate mesenchymal cells, *J Cell Physiol,* 177, 36-46.

Oates, P. J. and Mylari, B. L. (1999). Aldose reductase inhibitors: therapeutic implications for diabetic complications, *Expert Opin Investig Drugs,* 8, 2095-2119.

Oh, Y., Nagalla, S. R., Yamanaka, Y., Kim, H. S., Wilson, E. and Rosenfeld, R. G. (1996). Synthesis and characterization of insulin-like growth factor-binding protein (IGFBP)-7. Recombinant human mac25 protein specifically binds IGF-I and -II, *J Biol Chem,* 271, 30322-5.

Osterby, R., Parving, H. H., Hommel, E., Jorgensen, H. E. and Lokkegaard, H. (1990). Glomerular structure and function in diabetic nephropathy. Early to advanced stages, *Diabetes,* 39, 1057-63.

Pabst, R. and Sterzel, R. B. (1983). Cell renewal of glomerular cell types in normal rats. An autoradiographic analysis, *Kidney Int,* 24, 626-31.

Page, R., Morris, C., Williams, J., von Ruhland, C. and Malik, A. N. (1997). Isolation of diabetes-associated kidney genes using differential display, *Biochem Biophys Res Commun,* 232, 49-53.

Palmberg, P., Smith, M., Waltman, S., Krupin, T., Singer, P., Burgess, D., Wendtlant, T., Achtenberg, J., Cryer, P., Santiago, J., *et al.* (1981). The natural history of retinopathy in insulin-dependent juvenile-onset diabetes, *Ophthalmology,* 88, 613-8.

Parving, H. H., Smidt, U. M., Friisberg, B., Bonnevie-Nielsen, V. and Andersen, A. R. (1981). A prospective study of glomerular filtration rate and arterial blood pressure in insulin-dependent diabetics with diabetic nephropathy, *Diabetologia,* 20, 457-61.

Peng, C. and Mukai, S. T. (2000). Activins and their receptors in female reproduction, *Biochem Cell Biol,* 78, 261-79.

Persson, B. (2000). Bioinformatics in protein analysis, *Exs,* 88, 215-31.

Pichler, R. H., Bassuk, J. A., Hugo, C., Reed, M. J., Eng, E., Gordon, K. L., Pippin, J., Alpers, C. E., Couser, W. G., Sage, E. H., *et al.* (1996). SPARC is expressed by mesangial cells in experimental mesangial proliferative nephritis and inhibits platelet-derived-growth-factor-medicated mesangial cell proliferation in vitro, *Am J Pathol,* 148, 1153-67.

Pirart, J., Lauvaux, J. P. and Rey, W. (1978). Blood sugar and diabetic complications, *N Engl J Med,* 298, 1149.

Polyak, K., Lee, M. H., Erdjument-Bromage, H., Koff, A., Roberts, J. M., Tempst, P. and Massague, J. (1994). Cloning of p27Kip1, a cyclin-dependent kinase inhibitor and a potential mediator of extracellular antimitogenic signals, *Cell,* 78, 59-66.

Polyak, K., Xia, Y., Zweier, J. L., Kinzler, K. W. and Vogelstein, B. (1997). A model for p53-induced apoptosis see comments, *Nature,* 389, 300-5.

Pugliese, G., Pricci, F., Mene, P., Romeo, G., Nofroni, I., Giannini, S., Cresci, B., Galli, G., Rotella, C. M., Di Mario, U., *et al.* (1997). High glucose level unmasks a genetic predisposition to enhanced extracellular matrix production in mesangial cells from the Milan normotensive strain, *J Am Soc Nephrol,* 8, 406-14.

Pyronnet, S., Pradayrol, L. and Sonenberg, N. (2000). A cell cycle-dependent internal ribosome entry site, *Mol Cell,* 5, 607-16.

Rabinovich, G. A., Riera, C. M. and Sotomayor, C. E. (1999). Galectin-1, an alternative signal for T cell death, is increased in activated macrophages, *Braz J Med Biol Res,* 32, 557-67.

Rabinovich, G. A., Sotomayor, C. E., Riera, C. M., Bianco, I. and Correa, S. G. (2000). Evidence of a role for galectin-1 in acute inflammation, *Eur J Immunol,* 30, 1331-9.

Rana, B. K., Pan, L. and Insel, P. A. (2001). Use of an in silico approach to define the gene structure of eukaryotic adenylyl cyclases, *Biochem Biophys Res Commun,* 285, 152-7.

Risdon, R. A. (1985) In *Postgraduate Nephrology*, Vol. 1 (Ed, Marsh, F.) William Heinemann Medical Books Ltd, London, pp. 1-20.

Riser, B. L., Cortes, P., Heilig, C., Grondin, J., Ladson-Wofford, S., Patterson, D. and Narins, R. G. (1996). Cyclic stretching force selectively up-regulates transforming growth factor-beta isoforms in cultured rat mesangial cells, *Am J Pathol,* 148, 1915-23.

Riser, B. L., Ladson-Wofford, S., Sharba, A., Cortes, P., Drake, K., Guerin, C. J., Yee, J., Choi, M. E., Segarini, P. R. and Narins, R. G. (1999). TGF-beta receptor expression and binding in rat mesangial cells: modulation by glucose and cyclic mechanical strain, *Kidney Int,* 56, 428-39.

Riser, B. L., Varani, J., Cortes, P., Yee, J., Dame, M. and Sharba, A. K. (2001). Cyclic stretching of mesangial cells up-regulates intercellular adhesion molecule-1

and leukocyte adherence: a possible new mechanism for glomerulosclerosis, *Am J Pathol,* 158, 11-7.

Rogaev, E. I., Sherrington, R., Wu, C., Levesque, G., Liang, Y., Rogaeva, E. A., Ikeda, M., Holman, K., Lin, C., Lukiw, W. J., *et al.* (1997). Analysis of the 5' sequence, genomic structure, and alternative splicing of the presenilin-1 gene (PSEN1) associated with early onset Alzheimer disease, *Genomics,* 40, 415-24.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., *et al.* (2000). Comparative genomics of the eukaryotes, *Science,* 287, 2204-15.

Rupprecht, H. D., Schocklmann, H. O. and Sterzel, R. B. (1996). Cell-matrix interactions in the glomerular mesangium, *Kidney Int,* 49, 1575-82.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature,* 409, 928-33.

Saito, K., Shimizu, F., Sato, T. and Oite, T. (1993). Modulation of human mesangial cell behaviour by extracellular matrix components--the possible role of interstitial type III collagen, *Clin Exp Immunol,* 91, 510-5.

Sanai, T., Sobka, T., Johnson, T., el-Essawy, M., Muchaneta-Kubara, E. C., Ben Gharbia, O., el Oldroyd, S. and Nahas, A. M. (2000). Expression of cytoskeletal proteins during the course of experimental diabetic nephropathy, *Diabetologia,* 43, 91-100.

Santa Cruz, D. J., Hamilton, P. D., Klos, D. J. and Fernandez-Pol, J. A. (1997). Differential expression of metallopanstimulin/S27 ribosomal protein in melanocytic lesions of the skin, *J Cutan Pathol,* 24, 533-42.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray see comments, *Science,* 270, 467-70.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., *et al.* (2000). A gene expression database for the molecular pharmacology of cancer see comments, *Nat Genet,* 24, 236-44.

Schleicher, E. D. and Weigert, C. (2000). Role of the hexosamine biosynthetic pathway in diabetic nephropathy, *Kidney Int,* 58 Suppl 77, S13-8.

Schmidt, S., Schone, N. and Ritz, E. (1995). Association of ACE gene polymorphism and diabetic nephropathy? The Diabetic Nephropathy Study Group published erratum appears in Kidney Int 1995 Sep;48(3):915, *Kidney Int,* 47, 1176-81.

Schmitt, A. O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C. P., Hinzmann, B. and Rosenthal, A. (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues, *Nucleic Acids Res,* 27, 4251-60.

Schnaper, H. W. (2000). Signal transduction through protein kinase C, *Pediatr Nephrol,* 14, 254-8.

Schocklmann, H. O., Lang, S., Kralewski, M., Hartner, A., Ludke, A. and Sterzel, R. B. (2000). Distinct structural forms of type I collagen modulate cell cycle regulatory proteins in mesangial cells, *Kidney Int,* 58, 1108-20.

Schocklmann, H. O., Lang, S. and Sterzel, R. B. (1999). Regulation of mesangial cell proliferation, *Kidney Int,* 56, 1199-207.

Schreiner, G. F., Kiely, J. M., Cotran, R. S. and Unanue, E. R. (1981). Characterization of resident glomerular cells in the rat expressing Ia determinants and manifesting genetically restricted interactions with lymphocytes, *J Clin Invest,* 68, 920-31.

Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., *et al.* (1996). A gene map of the human genome, *Science,* 274, 540-6.

Sewell, W. A., Palmer, R. W., Spurr, N. K., Sheer, D., Brown, M. H., Bell, Y. and Crumpton, M. J. (1988). The human LFA-3 gene is located at the same chromosome band as the gene for its receptor CD2, *Immunogenetics,* 28, 278-82.

Shankland, S. J., Ly, H., Thai, K. and Scholey, J. W. (1996). Glomerular expression of tissue inhibitor of metalloproteinase (TIMP-1) in normal and diabetic rats, *J Am Soc Nephrol,* 7, 97-104.

Shapland, C., Hsuan, J. J., Totty, N. F. and Lawson, D. (1993). Purification and properties of transgelin: a transformation and shape change sensitive actin-gelling protein, *J Cell Biol,* 121, 1065-73.

Shapland, C., Lowings, P. and Lawson, D. (1988). Identification of new actin-associated polypeptides that are modified by viral transformation and changes in cell shape, *J Cell Biol,* 107, 153-61.

Sherlock, G. (2000). Analysis of large-scale gene expression data, *Curr Opin Immunol,* 12, 201-5.

Shore, A. C. and Tooke, J. E. (1994). Microvascular function in human essential hypertension, *J Hypertens,* 12, 717-28.

Simonson, M. S., Culp, L. A. and Dunn, M. J. (1989). Rat mesangial cell-matrix interactions in culture, *Exp Cell Res,* 184, 484-98.

Snedecor, G. W. and Cochran, W. G. (1973) In *Statistical Methods*Iowa State University Press, Ames, Iowa. USA, pp. Chapter 7.

Sorek, R. and Safer, H. M. (2003). A novel algorithm for computational identification of contaminated EST libraries, *Nucleic Acids Res,* 31, 1067-74.

Spillmann, A. A., Bandtlow, C. E., Lottspeich, F., Keller, F. and Schwab, M. E. (1998). Identification and characterization of a bovine neurite growth inhibitor (bNI-220), *J Biol Chem,* 273, 19283-93.

Sraer, J. D., Delarue, F., Hagege, J., Feunteun, J., Pinet, F., Nguyen, G. and Rondeau, E. (1996). Stable cell lines of T-SV40 immortalized human glomerular mesangial cells, *Kidney Int,* 49, 267-70.

Stamm, S. (2002). Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome, *Hum Mol Genet,* 11, 2409-16.

Stanier, P., Abu-Hayyeh, S., Murdoch, J. N., Eddleston, J. and Copp, A. J. (1998). Paralogous sm22alpha (Tagln) genes map to mouse chromosomes 1 and 9: further evidence for a paralogous relationship, *Genomics,* 51, 144-7.

Steffes, M. W., Osterby, R., Chavers, B. and Mauer, S. M. (1989). Mesangial expansion as a central mechanism for loss of kidney function in diabetic patients, *Diabetes,* 38, 1077-81.

Stephenson, L. A., Haney, L. B., Hussaini, I. M., Karns, L. R. and Glass, W. F., 2nd (1998). Regulation of smooth muscle alpha-actin expression and hypertrophy in cultured mesangial cells, *Kidney Int,* 54, 1175-87.

Sterzel, R. B., Schulze-Lohoff, E., Weber, M. and Goodman, S. L. (1992). Interactions between glomerular mesangial cells, cytokines, and extracellular matrix, *J Am Soc Nephrol,* 2, S126-31.

Stockand, J. D. and Sansom, S. C. (1998). Glomerular mesangial cells: electrophysiology and regulation of contraction, *Physiol Rev,* 78, 723-44.

Stollberg, J., Urschitz, J., Urban, Z. and Boyd, C. D. (2000). A quantitative evaluation of SAGE, *Genome Res,* 10, 1241-8.

Stoneking, M. (2001). Single nucleotide polymorphisms. From the evolutionary past, *Nature,* 409, 821-2.

Striker, L. J., Peten, E. P., Elliot, S. J., Doi, T. and Striker, G. E. (1991). Mesangial cell turnover: effect of heparin and peptide growth factors, *Lab Invest,* 64, 446-56.

Studer, R. K., Negrete, H., Craven, P. A. and DeRubertis, F. R. (1995). Protein kinase C signals thromboxane induced increases in fibronectin synthesis and TGF-beta bioactivity in mesangial cells, *Kidney Int,* 48, 422-30.

Sugiyama, H., Kashihara, N., Maeshima, Y., Okamoto, K., Kanao, K., Sekikawa, T. and Makino, H. (1998). Regulation of survival and death of mesangial cells by extracellular matrix, *Kidney Int,* 54, 1188-96.

Suzuki, D. and Miyata, T. (1999). Carbonyl stress in the pathogenesis of diabetic nephropathy, *Intern Med,* 38, 309-14.

Suzuki, T., Hashimoto, S., Toyoda, N., Nagai, S., Yamazaki, N., Dong, H. Y., Sakai, J., Yamashita, T., Nukiwa, T. and Matsushima, K. (2000). Comprehensive gene expression profile of LPS-stimulated human monocytes by SAGE, *Blood,* 96, 2584-91.

Tang, B. L., Low, D. Y. and Hong, W. (1998). Hsec22c: a homolog of yeast Sec22p and mammalian rsec22a and msec22b/ERS-24, *Biochem Biophys Res Commun,* 243, 885-91.

Tisher, C. C. and Madsen, K. M. (2000) In *Brenner & Rectors The Kidney*, Vol. 1 (Ed, Brenner, B. M.) W. B. Saunders, Philadelphia, pp. 3-67.

Tomiuk, S. and Hofmann, K. (2001). Microarray probe selection strategies, *Brief Bioinform,* 2, 329-40.

Tomlinson, D. R. (1999). Mitogen-activated protein kinases as glucose transducers for diabetic complications, *Diabetologia,* 42, 1271-81.

Trachtman, H. (1994). Vitamin E prevents glucose-induced lipid peroxidation and increased collagen production in cultured rat mesangial cells, *Microvasc Res,* 47, 232-9.

Trachtman, H., Futterweit, S. and Bienkowski, R. S. (1993). Taurine prevents glucose-induced lipid peroxidation and increased collagen production in cultured rat mesangial cells, *Biochem Biophys Res Commun,* 191, 759-65.

Tremain, N., Korkko, J., Ibberson, D., Kopen, G. C., DiGirolamo, C. and Phinney, D. G. (2001). MicroSAGE Analysis of 2,353 Expressed Genes in a Single Cell-Derived Colony of Undifferentiated Human Mesenchymal Stem Cells Reveals mRNAs of Multiple Cell Lineages, *Stem Cells,* 19, 408-18.

Trevison, R., Walker, J. D. and Viberti, G. (1997) In *Nephrology*(Eds, Jamison, R. and Wilkinson, R.) Chapman & Hall, London, pp. 551-574.

Tsuchiya, S., Kobayashi, Y., Goto, Y., Okumura, H., Nakae, S., Konno, T. and Tada, K. (1982). Induction of maturation in cultured human monocytic leukemia cells by a phorbol diester, *Cancer Res,* 42, 1530-6.

Tsuchiya, S., Yamabe, M., Yamaguchi, Y., Kobayashi, Y., Konno, T. and Tada, K. (1980). Establishment and characterization of a human acute monocytic leukemia cell line (THP-1), *Int J Cancer,* 26, 171-6.

Tsui, S. K., Lee, S. M., Fung, K. P., Waye, M. M. and Lee, C. Y. (1996). Primary structures and sequence analysis of human ribosomal proteins L39 and S27, *Biochem Mol Biol Int,* 40, 611-6.

Velculescu, V. E., Vogelstein, B. and Kinzler, K. W. (2000). Analysing uncharted transcriptomes with SAGE, *Trends Genet,* 16, 423-5.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995). Serial analysis of gene expression, *Science,* 270, 484-7.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B. and Kinzler, K. W. (1997). Characterization of the yeast transcriptome, *Cell,* 88, 243-51.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome, *Science,* 291, 1304-51.

Virlon, B., Cheval, L., Buhler, J. M., Billon, E., Doucet, A. and Elalouf, J. M. (1999). Serial microanalysis of renal transcriptomes, *Proc Natl Acad Sci U S A,* 96, 15286-91.

Wada, J., Zhang, H., Tsuchiyama, Y., Hiragushi, K., Hida, K., Shikata, K., Kanwar, Y. S. and Makino, H. (2001). Gene expression profile in streptozotocin-induced diabetic mice kidneys undergoing glomerulosclerosis, *Kidney Int,* 59, 1363-73.

Wahab, N. A., Parker, S., Sraer, J. D. and Mason, R. M. (2000). The decorin high glucose response element and mechanism of its activation in human mesangial cells, *J Am Soc Nephrol,* 11, 1607-19.

Welle, S., Bhatt, K. and Thornton, C. A. (1999). Inventory of high-abundance mRNAs in skeletal muscle of normal men, *Genome Res,* 9, 506-13.

Williams, B., Gallacher, B., Patel, H. and Orme, C. (1997). Glucose-induced protein kinase C activation regulates vascular permeability factor mRNA expression and peptide production by human vascular smooth muscle cells in vitro, *Diabetes,* 46, 1497-503.

Williamson, J. R., Chang, K., Frangos, M., Hasan, K. S., Ido, Y., Kawamura, T., Nyengaard, J. R., van den Enden, M., Kilo, C. and Tilton, R. G. (1993). Hyperglycemic pseudohypoxia and diabetic complications, *Diabetes,* 42, 801-13.

Wilmer, W. A. and Cosio, F. G. (1998). DNA binding of activator protein-1 is increased in human mesangial cells cultured in high glucose concentrations, *Kidney Int,* 53, 1172-81.

Wogensen, L., Nielsen, C. B., Hjorth, P., Rasmussen, L. M., Nielsen, A. H., Gross, K., Sarvetnick, N. and Ledet, T. (1999). Under control of the Ren-1c promoter, locally produced transforming growth factor-beta1 induces accumulation of glomerular extracellular matrix in transgenic mice, *Diabetes,* 48, 182-92.

Wolf, B. A., Williamson, J. R., Easom, R. A., Chang, K., Sherman, W. R. and Turk, J. (1991). Diacylglycerol accumulation and microvascular abnormalities induced by elevated glucose levels, *J Clin Invest,* 87, 31-8.

Wolf, G., Schroeder, R., Thaiss, F., Ziyadeh, F. N., Helmchen, U. and Stahl, R. A. (1998). Glomerular expression of p27Kip1 in diabetic db/db mouse: role of hyperglycemia, *Kidney Int,* 53, 869-79.

Wolf, G., Schroeder, R., Ziyadeh, F. N., Thaiss, F., Zahner, G. and Stahl, R. A. (1997). High glucose stimulates expression of p27Kip1 in cultured mouse mesangial cells: relationship to hypertrophy, *Am J Physiol,* 273, F348-56.

Wolf, G. and Ziyadeh, F. N. (1999). Molecular mechanisms of diabetic renal hypertrophy, *Kidney Int,* 56, 393-405.

Xia, P., Inoguchi, T., Kern, T. S., Engerman, R. L., Oates, P. J. and King, G. L. (1994). Characterization of the mechanism for the chronic activation of diacylglycerol-protein kinase C pathway in diabetes and hypergalactosemia, *Diabetes,* 43, 1122-9.

Yagame, M., Kim, Y., Zhu, D., Suzuki, D., Eguchi, K., Nomoto, Y., Sakai, H., Groppoli, T., Steffes, M. W. and Mauer, S. M. (1995). Differential distribution of type IV collagen chains in patients with diabetic nephropathy in non-insulin-dependent diabetes mellitus, *Nephron,* 70, 42-8.

Yamamoto, T., Nakamura, T., Noble, N. A., Ruoslahti, E. and Border, W. A. (1993). Expression of transforming growth factor beta is elevated in human and experimental diabetic nephropathy, *Proc Natl Acad Sci U S A,* 90, 1814-8.

Yan, S. D., Schmidt, A. M., Anderson, G. M., Zhang, J., Brett, J., Zou, Y. S., Pinsky, D. and Stern, D. (1994). Enhanced cellular oxidant stress by the interaction of advanced glycation end products with their receptors/binding proteins, *J Biol Chem,* 269, 9889-97.

Yano, N., Endoh, M., Fadden, K., Yamashita, H., Kane, A., Sakai, H. and Rifai, A. (2000). Comprehensive gene expression profile of the adult human renal cortex: analysis by cDNA array hybridization, *Kidney Int,* 57, 1452-9.

Zaucke, F., Dinser, R., Maurer, P. and Paulsson, M. (2001). Cartilage oligomeric matrix protein (COMP) and collagen IX are sensitive markers for the differentiation state of articular primary chondrocytes, *Biochem J,* 358, 17-24.

Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. and Kinzler, K. W. (1997). Gene expression profiles in normal and cancer cells, *Science,* 276, 1268-72.

Zheng, P., Eastman, J., Vande Pol, S. and Pimplikar, S. W. (1998). PAT1, a microtubule-interacting protein, recognizes the basolateral sorting signal of amyloid precursor protein, *Proc Natl Acad Sci U S A,* 95, 14745-50.

Ziyadeh, F. N., Han, D. C., Cohen, J. A., Guo, J. and Cohen, M. P. (1998). Glycated albumin stimulates fibronectin gene expression in glomerular mesangial cells: involvement of the transforming growth factor-beta system, *Kidney Int,* 53, 631-8.

Ziyadeh, F. N., Hoffman, B. B., Han, D. C., Iglesias-De La Cruz, M. C., Hong, S. W., Isono, M., Chen, S., McGowan, T. A. and Sharma, K. (2000). Long-term prevention of renal insufficiency, excess matrix gene expression, and glomerular

mesangial matrix expansion by treatment with monoclonal antitransforming growth factor-beta antibody in db/db diabetic mice see comments, *Proc Natl Acad Sci U S A,* 97, 8015-20.

APPENDIX 1

GENE ABBREVIATIONS

Alphabetical by abbreviation

Abbreviation	HGNC (Official gene symbol & Name)	UniGene Cluster ID	Gene Description
a2-SMA	ACTA2	Hs.195851	Actin, alpha 2, smooth muscle
ACTA	INHBA	Hs.727	Inhibin, beta A (activin A)
ADP/ATP	SLC25A6	Hs.164280	Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6
AR	AKR1B1	Hs.75313	Aldo-keto reductase family 1, member B1 (aldose reductase)
BACT	ACTB	Hs.180952 (Hs.288062)	Actin, beta
BMP5	BMP5	Hs.1104	Bone morphogenetic protein 5
BSAP-1	BASP1	Hs.79516	Brain abundant, membrane attached signal protein 1
CALM3	CALM3	Hs.334330	Calmodulin 3 (phosphorylase kinase, delta)
CCND1	CCND1	Hs.82932	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)
CD151	CD151	Hs.75564	CD151 antigen
COLL Ia2	COL1A2	Hs.179573	Collagen, type I, alpha 2
COLL IIIa1	COL3A1	Hs.119571	Collagen, type III, alpha 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)
COLL IVa1	COL4A1	Hs.119129	Collagen, type IV, alpha 1
COLL IVa2	COL4A2	Hs.75617	Collagen, type IV, alpha 2
COX6a1	COX6A1	Hs.180714	Cytochrome c oxidase subunit VIa polypeptide 1
CR1	DCXR	Hs.9857	Dicarbonyl/L-xylulose reductase
CRYAB	CRYAB	Hs.1940	Crystallin, alpha B
CSPG4	CSPG4	Hs.9004	Chondroitin sulfate proteoglycan 4 (melanoma-associated)
CSTB	CSTB	Hs.695	Cystatin B (stefin B)
CTGF	CTGF	Hs.75511	Connective tissue growth factor
DAXX	MLC-B	Hs.180224	Myosin regulatory light chain
DBI	DBI	Hs.78888	Diazepam binding inhibitor (GABA receptor modulator, acyl-Coenzyme A binding protein)
DCIP1	CCNDBP1	Hs.3674 (Hs.36794)	Cyclin D-type binding-protein 1
ECM1	ECM1	Hs.81071	Extracellular matrix protein 1
EF1a	EEF1A1	Hs.181165	Eukaryotic translation elongation factor 1 alpha 1
EF2	EEF2	Hs.75309	Eukaryotic translation elongation factor 2
ENIGMA	ENIGMA	Hs.102948	Enigma (LIM domain protein)
ENSa	ENSA	Hs.111680	Endosulfine alpha
EPLIN	EPLIN	Hs.10706	Epithelial protein lost in neoplasm beta
FBN-1	FBN1	Hs.750	Fibrillin 1 (Marfan syndrome)
FK506bp9	FKBP9	Hs.302749	FK506 binding protein 9, 63 kDa
FN-1	FN1	Hs.287820	Fibronectin 1
GAL1	LGALS1	Hs.227751	Lectin, galactoside-binding, soluble, 1 (galectin 1)
GAPDH	GAPD	Hs.169486 (Hs.169476)	Glyceraldehyde-3-phosphate dehydrogenase
GFAT	GFPT1	Hs.1674	Glutamine-fructose-6-phosphate transaminase 1
GLA	MGP	Hs.75742 (Hs.365706)	Matrix Gla protein

Abbreviation	HGNC (Official gene symbol & Name)	UniGene Cluster ID	Gene Description
GLO1	GLO1	Hs.75207	Glyoxalase I
GPX1	GPX1	Hs.76686	Glutathione peroxidase 1
GRN	GRN	Hs.180577	Granulin
HMG1	HMGB1	Hs.337757	High-mobility group box 1
HSP27	HSPB1	Hs.76067	Heat shock 27kDa protein 1
HSP70	HSPA5	Hs.75410	Heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)
IGFbp7	IGFBP7	Hs.119206	Insulin-like growth factor binding protein 7
LRP-1		Hs.475478	LDL receptor related protein-1
LTGFbp4	LTBP4	Hs.85087	Latent transforming growth factor beta binding protein 4
MAPKPase	DUSP16	Hs.20281	Dual specificity phosphatase 16
MCT-1	MCT-1	Hs.102696	MCT-1 protein
MFGE8	MFGE8	Hs.3745	Milk fat globule-EGF factor 8 protein
MGSH	MGST1	Hs.790 (Hs.355733)	Microsomal glutathione S-transferase 1
MLC-B	MLC-B	Hs.233936	Myosin regulatory light chain
NDUF1	NDUFA1	Hs.74823	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa
OXYrc	OXTR	Hs.2820	Oxytocin receptor
p27 kip1	CDKN1B	Hs.238990	Cyclin-dependent kinase inhibitor 1B (p27, Kip1)
P4HB	P4HB	Hs.387107(Hs. 75655)	Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), beta
PGM1	PGM1	Hs.1869	Phosphoglucomutase 1
PKC b1	PRKCB1	Hs.77202	Protein kinase C, beta 1
PKC i	PRKCI	Hs.1904	Protein kinase C, iota
PKC n	PRKCN	Hs.143460	Protein kinase C, nu
PKC z	PRKCZ	Hs.78793	Protein kinase C, zeta
PKCμ	PRKCM	Hs.2891	Protein kinase C, mu
QSCN6	QSCN6	Hs.77266	Quiesin Q6
RP			Ribsomal Proteins
RTN4	RTN4	Hs.65450	Reticulon 4
SERPINb6	SERPINB6	Hs.41072	Serine (or cysteine) proteinase inhibitor, clade B, member 6
SPARC	SPARC	Hs.111779	Secreted protein, acidic, cysteine-rich (osteonectin)
TAGLN	TAGLN	Hs.75777	Transgelin
TC21	RRAS2	Hs.206097	Related RAS viral (r-ras) oncogene, homologue 2.
TEF-3	TCF3	Hs.101047	Transcription factor 3
TEF-A	TCEAL1	Hs.95243	Transcription elongation factor A (SII)-like 1
TGF b1	TGFB1	Hs.1103	Transforming growth factor, beta 1
TGF b2	TGFB2	Hs.169300	Transforming growth factor, beta 2
TFG b3	TGFB3	Hs.2025	Transforming growth factor, beta 3
TRAILrc	TNFRSF10B	Hs.51233	Tumor necrosis factor receptor superfamily, member 10b
TSP1	THBS1	Hs.87409	Thrombospondin 1
VIM	VIM	Hs.2064 (Hs.297753)	Vimentin
X11-like	APBA2	Hs.26468	Amyloid beta (A4) precursor protein-binding, family A, member 2 (X11-like)

APPENDIX 2

PRIMER SEQUENCES & REFERENCE ACCESSION NUMBERS

Aplhabetical by gene abbreviation

Gene	Platform		Primers	Amplicon	Ref Seq Acc
A2SMA	ABI	For	GCACCCAGCACCATGAAGAT	66	NM001613
		Rev	ACCGATCCAGACAGAGTATTTGC		
A2SMA.lc	LC	For	AAGGCCAACCGGGAGAAAATGACT	467	NM001613
		Rev	CCGATGAAGGATGGCTGGAACA		
ACTA	ABI	For	CAAGCTGAGACCCATGTCCAT	75	NM002192
		Rev	TCTGAATGTCCTTTTTGATGATGTTT		
ACTA.lc	LC	For	TCGGGGAGAACGGGTATGTGGAGA	407	NM002192
		Rev	CAACCGCTGGATGCTGCTGGAGA		
ADPtr	ABI	For	CCATGGGTGAGACACTCCAGTA	61	J03592
		Rev	CGACTTGGCTCCTACAAGCAT		
APPBP2	ABI	For	CCGGGTTCCAGCGATTCT	57	NM006380
		Rev	GGAGCCTGTAATCCCAGCTACTC		
bACT.lc	LC	For	ATGATATCGCCGCGCTCGTCGTC	547	BC009275
		Rev	AGGTCCCGGCCAGCCAGGTCCAG		
CALM3	ABI	For	CCCAGCGGAGAGCATGAT	56	NM005184
		Rev	ACTTGTCCCAGAGGCGAAAGT		
CD151	ABI	For	ACTGTGCACTGCCCTGTTCA	76	NM004357
		Rev	CAAAACCAGGAAGGCCCTG		
CD151.lc	LC	For	TGGCGGGCACTGTCGTCAT	428	NM004357
		Rev	GCAGCCGCCCTCCACCTTGTAGAT		
CD58	ABI	For	TTCTTTCTTTATGTGCTTGAGTCTCTTC	133	NM001779
		Rev	CCCATGAGTACATTATAAGTCCTCGAT		
CKB	ABI	For	CCATGCACCCCTCGATGT	57	NM001823
		Rev	GCCTTCCTTACAGCAAGGCTAA		
COLL Ia2	ABI	For	ATGCCAGCTCCATGTTCTCC	62	NM000089
		Rev	GGCAAAAGAATGACCTACCGC		
COLLIIIa1	ABI	For	CACAGGGATTCTCCTCCTTCAT	91	NM000090
		Rev	AACACATTCTCTATGCTAGTGTGACAAA		
COLLIIIa1.lc	LC	For	GGGGTCCTACTGGTCCTATTGG	400	NM000090
		Rev	CCTGGGGGTCCTGGGTTAC		
COLLIVa1	ABI	For	GACCGCAGGAGGGCAGAT	62	NM001845
		Rev	TTGGAGACTTTTAGTGAAATGTCATTTC		
COLLIVa1.lc	LC	For	CCTGCCGGGCCTACTGGT	450	NM001845
		Rev	GGCACGGTGGGATCTGAATGGT		
COX6a1	ABI	For	CAGCACTGGTTTGGACCGTTA	73	NM004373
		Rev	TAAAGAAGGTTAGCTTAAGGTCCCATA		
CR	ABI	For	CCATGCCGTGCTCATCCT	57	NM016286
		Rev	TTGGGCAGCAGAATCAGGTT		
CSPG4	ABI	For	TTGATGGGCCAAGGGCTAA	80	NM001897
		Rev	GGCCTGAGACCCTCGATGA		
CSPG4	LC	For	CATCCTGCCCCTGCTCTTCTACCT	418	NM001897
		Rev	CCGGACCCCTGGGACTATCTC		
CSTB	ABI	For	ACCAACAAAGCCAAGCATGAT	59	NM000100
		Rev	GGCCTTGTCCAAAGTCAGGAT		
CTGF	ABI	For	TGATATGACTGTTTTTCGGACAGTTTA	75	NM001901
		Rev	CAACTAGAAAGGTGCAAACATGTAACTT		
CTGF2	ABI	For	TGGAAATTCTCTCAGATAGAATGACAGT	64	NM001901
		Rev	TGATGCCTCCCCTTTGCA		
CTGF.lc	LC	For	ACGGCGAGGTCATGAAGAAGAACA	523	NM001901
		Rev	TGGGGCTACAGGCAGGTCAGTG		

Gene	Platform		Primers	Amplicon	Ref Seq Acc
DCIP-1	ABI	For	CGTCTCCACAGGAAACCCAG	60	NM012142
		Rev	TGATGGCAGCATGGACTTGT		
DCIP-1	LC	For	AGAGCCCTGAGAACAATGACCTTA	396	NM012142
		Rev	TCTGTGCCACCTGATCCTTCTTC		
ENSa	ABI	For	CACGCAGGAGAAAGAAGGTATTC	65	NM004436
		Rev	GGTATTTGGCCTTTAGCTTTGC		
EST1	ABI	For	CAGTATTTAACATTCCCCCAAAGAA	68	AA401448
		Rev	GTAATATCACAGTATGGGACAAAGGTTT		
EST2	ABI	For	TCGACAGAGTAAGGCCCATCTC	74	W77990
		Rev	GCGCTGGCGGATGCT		
EST3	ABI	For	TGCCAGTTCTCTTGTGTGCAA	65	H94936
		Rev	ATCCTGGCACAGGGTGTTAGA		
EST4	ABI	For	CAGACCTCATTATATGCTTTCATGATTC	68	R23924
		Rev	GCATTAGTGGTCTGATTGGAAAGA		
EST5	ABI	For	TCTCTTGGGTCCCTTCCATGT	55	R99311
		Rev	AGCCAGCATGCCCCATT		
FBN1	ABI	For	GCATGTGCAATATGCCAAGATT	63	NM000138
		Rev	TTTATGACATTGACCCCTTGTTGA		
FBN1.lc	LC	For	TCCCGGATTTACCCAACACCATAC	435	NM000138
		Rev	GCCTGCGCAGAGCCACATT		
FK506BP9	LC	For	CAGCCTCCGCCTCCCGTATTCA	317	BC007443
		Rev	GCCCAGCCCCCAGCCCTATTTAT		
FN1	ABI	For	TGGCATTGCCAACCTTTACA	51	X02761
		Rev	TTCGACAGGACCACTTGAGCT		
GAL1.lc	LC	For	AATCATGGCTTGTGGTCTGG	428	BC001693
		Rev	CATGGGCTGGCTGATTTC		
GAPDH	ABI	For	TTGTCAAGCTCATTTCCTGGTATG	60	NM002046
		Rev	GGTCCACCACCCTGTTGCT		
GAPDH.lc	LC	For	GCGGGGCTCTCCAGAACATCATCC	470	NM002046
		Rev	TGCCAGCCCCAGCGTCAAAGGTG		
GLA.lc	LC	For	TGACCTGCAGGACGAAACC	400	BC005272
		Rev	TGCTACAGGGGGATACAAAAT		
GLO1	ABI	For	GCCATGATTCACATTTGATGAGTT	151	NM006708
		Rev	CGGTTGGCATGGCCTTT		
GPX1	ABI	For	GAGATTCTGAATTCCCTCAAGTACGT	73	NM000581
		Rev	TTCACCTCGCACTTCTCGAA		
HMG1	ABI	For	CCACTAACCTTGCCTGGTACAGTAT	73	NM002128
		Rev	GCACCAACAAGAACCTGCTTTAA		
HSP70	ABI	For	ACCTGGGTTAGGGTGTGTGTTC	82	NM005347
		Rev	GAAAAAACTTCCTACACCAGATGCA		
HSP70.lc	LC	For	AAGAGGGAGGGGGAGAAGAACATC	463	NM005347
		Rev	GAGTCGAGCCACCAACAAGAACAA		
IGFbp7.lc	LC	For	CGTGAAGAGCCGCAAGAGG	448	L19182
		Rev	AGATACCAGCACCCAGCCAGTTAC		
MCT-1	LC	For	GCGCGGCTGGCTCTCGT	435	BC001013
		Rev	TGGTGTGGCAGGATAAAAGGATA		
mGSH	ABI	For	CATGTTATGATTTGTAACATTCACACAAC	76	BC005923
		Rev	ATAGGTTTCTCATACGTGCAATTCTTT		
NADH	ABI	For	CAACGGAGGAGGCTACTACCACTA	74	NM004541
		Rev	AGATGTCCACGGGCACGAT		
NDUFa1	ABI	For	CGCATCTCTGGAGTTGATCGT	151	NM004541
		Rev	CACATTTGCATGCTACATAATACACTGT		
OXYrc	ABI	For	GGCTGTGAGAGATGAGGCATG	55	NM000916
		Rev	TCACCCGAAAAAGAAACCCC		
PGM1	ABI	For	AGAGGACCTGCGGGCTTAGA	50	NM002633
		Rev	GCAGGAGGGCATGAAAAGG		
PGM1	LC	For	GGCCACCACCCTGACCCCAACC	539	NM002633
		Rev	GCGCCTACGCTTCCACCTCCTC		

Gene	Platform		Primers	Amplicon	Ref Seq Acc
SEC22A	ABI	For	TGTCAGAGATGGACTGCCACTT	62	NM012430
		Rev	CCTGCATTCCTGTGCTTTGTT		
SERPIN b6	ABI	For	TGGGCAAGGCAGACTTCTCT	63	NM004568
		Rev	TGTGCACGACCTTGGACAGA		
SERPINb6	LC	For	ATCATGCTTCCGCACGAGACCACT	425	NM004568
		Rev	AAGCGGCCGCAGAAGAGAATCC		
SPARC	ABI	For	ATCTCACAGGCTGAGAACTCGTT	92	NM003118
		Rev	CTCTTCACATCATGGTGAGAGTTT		
SPARC.lc	LC	For	TGGGCAAAGGGAAGTAACAGACAC	446	NM003118
		Rev	CAACCGATTCACCAACTCCACTTT		
TAGLN	ABI	For	CCAGGCCGGCATGACA	63	NM003186
		Rev	TAGCCCTCTCCGCTCTAACTGA		
TAGLN,lc	LC	For	AAGAAAGCGCAGGAGCATAAGAGG	409	NM003186
		Rev	AGCCCAGGGAGGAGACAGTAGAGG		
TGFb1	ABI	For	AGGACCTCGGCTGGAAGTG	61	NM000660
		Rev	AGTTGGCATGGTAGCCCTTG		
TGFb1.lc	LC	For	CTGCCTCCTCCTGCCTGTCT	308	NM000660
		Rev	CCCGCCTGGCCTGAACTACT		
TGFb2	ABI	For	TGGAGCATGCCCGTATTTATG	51	NM003238
		Rev	AGGACCCTGCTGTGCTGAGT		
TGFb2.lc	LC	For	CTGCGTGTCCCAAGATTTAGAAC	450	NM003238

APPENDIX 3

THP-1 1°TRANSCRIPTOME

Descending Tag Frequency

Tag Sequence	PMA	THP1	Total Tags	Redundant Matches	UniGene ID (Hs.)	Gene Description
TTATCAAGTG	33	73	106	1		No Match
TGCCAAGGGT	41	63	104	1		No Match
ATAATACATA	37	38	75	1		No Match
TTATATAGTG	21	30	51	1		No Match
AGGCAGACAG	26	24	50	2		Greater than one match
ATGTCTCAAA	20	28	48	1	114057	"ESTs, Weakly similar to ALU7_HUMAN ALU SUBFAMILY SQ SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens]"
GAGCGTTTTG	16	21	37	1		No Match
GCCTCCAAGG	10	27	37	1	169476	glyceraldehyde-3-phosphate dehydrogenase
GATTCCGTGA	13	20	33	1	179779	ribosomal protein L37
CAAGGTGACA	14	19	33	3		Greater than one match
CCCTGGGTTC	21	9	30	1	111334	"ferritin, light polypeptide"
CACAAACGGT	13	15	28	1	195453	ribosomal protein S27 (metallopanstimulin 1)
AAAAAGTACC	15	13	28	1		No Match
CCAGAACAGA	8	20	28	3		Greater than one match
AGAGCGAAGT	9	18	27	1	246074	"ESTs, Weakly similar to YZA1_HUMAN HYPOTHETICAL PROTEIN [H.sapiens]"
GTGTTAACCA	11	14	25	2		Greater than one match
AAGGAAATGG	7	17	24	3		Greater than one match
ATACTGACAT	9	14	23	1		No Match
GCCTTTATGA	10	13	23	1	8768	hypothetical protein FLJ10849
GATACTTGGA	9	13	22	1		No Match
TAAAGAGGCC	9	13	22	1		No Match
CTAATAAAGC	11	10	21	1		No Match
CCATCCGTAA	10	10	20	1		No Match
TGTGTTGAGA	16	4	20	2		Greater than one match
GTGGCTCACA	6	14	20	6		Greater than one match
AAGGTGGAAG	15	4	19	1	163593	ribosomal protein L18a
GGGTTTTTAT	12	7	19	1	74497	nuclease sensitive element binding protein 1
GTGAAACTAA	11	8	19	1	106671	cleft lip and palate associated transmembrane protein 1
AAGATCAAGA	7	12	19	3		Greater than one match
AATCCTGTGG	10	8	18	1	178551	ribosomal protein L8
CGCCGCCGGC	11	7	18	1	182825	ribosomal protein L35
ACATCATAGA	9	9	18	1	182979	ribosomal protein L12
AACAATTTGG	3	15	18	1		No Match
AGGACAAATA	10	8	18	1	172458	iduronate 2-sulfatase (Hunter syndrome)
GAATAATAAA	6	11	17	1	90078	nucleotide-sugar transporter similar to C. elegans sqv-7
ACCAAAATCC	7	10	17	1	7953	HSPC041 protein
CTGAACATCT	5	11	16	1	75835	phosphomannomutase 1
AGACAATAAC	5	11	16	1		No Match

Tag Sequence	PMA	THP1	Total Tags	Redundant Matches	UniGene ID (Hs.)	Gene Description
CCTACCAAGA	5	11	16	1		No Match
CTGTAGGTGA	10	6	16	1		No Match
GCCAAGTGGA	6	10	16	1		No Match
GAAGCAGGAC	9	7	16	1	180370	cofilin 1 (non-muscle)
GTCATAGCTG	7	9	16	3		Greater than one match
ATTACGCCAA	8	7	15	1		No Match
TGGTGTTGAG	11	3	14	1	275865	ribosomal protein S18
ATTGCTTAGA	8	6	14	1		No Match
CCCTGAGTCC	6	8	14	1		No Match
CCTTTAATCC	9	5	14	1		No Match
AGCGGATACA	4	9	13	1		No Match
GTCTGCTGAT	5	8	13	1		No Match
TCAGGCTGCC	6	7	13	1		No Match
GCAATCTGAT	5	8	13	1	290875	ESTs
GTATGGGCCC	8	5	13	1	75184	chitinase 3-like 1 (cartilage glycoprotein-39)
ATACTGAAGC	5	8	13	2		Greater than one match
CAGTCTCTCA	3	10	13	2		Greater than one match
TTGGTGAAGG	9	4	13	2		Greater than one match
AAAAAAAAAA	8	5	13	80		Greater than one match
TGGATCAGTC	8	4	12	1	91417	topoisomerase (DNA) II binding protein
AATATGTGTG	9	3	12	1		No Match
TGTGCCAAGT	7	5	12	1		No Match
TTTTATGTTT	3	9	12	1	15303	KIAA0349 protein
CCAAATAAAA	5	7	12	1	50842	interferon-induced protein 35
GGAAGCCACT	4	8	12	1	23213	ESTs
GGATGCTGGG	10	2	12	1	8438	ESTs
CCCCAGCCAG	7	4	11	1	252259	ribosomal protein S3
AGGTCGGGTG	5	6	11	1		No Match
ATGACTGATA	6	5	11	1		No Match
GGCTTCGGTC	6	5	11	1		No Match
GGGAAGGCGG	3	8	11	1		No Match
GTGAACGTGC	7	4	11	1		No Match
GTTGCTGAGA	4	7	11	1		No Match
GAAATTTAAA	6	5	11	1	274472	high-mobility group (nonhistone chromosomal) protein 1
CCACTGCACT	10	1	11	74		Greater than one match
GGCTTTGGTC	7	3	10	1	177592	"ribosomal protein, large, P1"
AAAACAGTGG	2	8	10	1	5566	ribosomal protein L37a
TTCATTATAA	5	5	10	1	250655	"prothymosin, alpha (gene sequence 28)"
GGCTGGGGGC	5	5	10	1	75721	profilin 1
TAAGATCCTT	8	2	10	1		No Match
TGAGCAAAAG	2	8	10	1		No Match
TGGGTTGTCT	2	8	10	1		No Match
ACAAACTTAG	6	4	10	1	177656	"calmodulin 1 (phosphorylase kinase, delta)"
AAGGTAGCAG	10	0	10	1	104125	adenylyl cyclase-associated protein
GTGACCACGG	9	1	10	2		Greater than one match
TGAAATAAAC	4	6	10	2		Greater than one match
GAAATGATGA	5	4	9	1	288856	prefoldin 5
CAAACCTCCA	5	4	9	1		No Match
CTATCAAGTG	4	5	9	1		No Match
GCCTAATGTA	3	6	9	1		No Match
CCTGATCTTT	5	4	9	1	181357	"laminin receptor 1 (67kD,

Tag Sequence	PMA	THP1	Total Tags	Redundant Matches	UniGene ID (Hs.)	Gene Description
						ribosomal protein SA)"
TCTACAAGAA	3	6	9	1	113994	"Homo sapiens cDNA FLJ20796 fis, clone COL00301"
TTGGGGTTTC	9	0	9	1	62954	"ferritin, heavy polypeptide 1"
TGGCCCCAGG	8	1	9	1	268571	apolipoprotein C-I
ATTCTCCAGT	4	5	9	2		Greater than one match
TTCTTTCCTG	2	7	9	2		Greater than one match
CCTGTAATCC	5	4	9	132		Greater than one match
CCCATCGTCC	5	3	8	1	mito	Tag matches mitochondrial sequence
CTGCTATCCG	3	5	8	1	180946	ribosomal protein L5
AAAAATCATC	5	3	8	1		No Match
CCCACAAGGT	3	5	8	1		No Match
CTAAGACTTC	8	0	8	1		No Match
CTAGTCTTTG	3	5	8	1		No Match
AAGAAAATAG	5	3	8	1	194657	"cadherin 1, type 1, E-cadherin (epithelial)"
CAGATCTTTG	3	5	8	2		Greater than one match
GGCAAGCCCC	6	2	8	2		Greater than one match
AGATCTATAC	3	5	8	3		Greater than one match
GCCGTGTCCG	6	1	7	1	241507	ribosomal protein S6
AGCTCTCCCT	5	2	7	1	82202	ribosomal protein L17
AGGCTACGGA	5	2	7	1	119122	ribosomal protein L13a
ATTTAGAGGT	6	1	7	1		No Match
CACCACCGTT	4	3	7	1		No Match
CACGGCTTTC	4	3	7	1		No Match
CTCGAGTCTC	2	5	7	1		No Match
GCCAAGGGTC	4	3	7	1		No Match
GTAAGCATAA	1	6	7	1		No Match
TAACTCGCCT	3	4	7	1		No Match
TCCCTGTGCG	3	4	7	1		No Match
TGGTGACAAA	3	4	7	1		No Match
TTCAGCTCGA	3	4	7	1		No Match
CACCACCACA	3	4	7	1	142856	KIAA1515 protein
TGGGCATCCA	3	4	7	1	19333	hypothetical protein FLJ10349
GCGGCGGATG	3	4	7	1	287389	"Human HL14 gene encoding beta-galactoside-binding lectin, 3' end, clone 2"
TGGGCAAAGC	6	1	7	1	2186	eukaryotic translation elongation factor 1 gamma
CTAGTGTTGA	3	4	7	1	147916	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 3
AGAAACAAGA	1	6	7	2		Greater than one match
GAGGCTTTGC	2	5	7	2		Greater than one match
CCTTTGAGAT	4	3	7	3		Greater than one match
TGGCGTACGG	5	1	6	1	ribo	Tag matches ribosomal RNA sequence
CCGTCCAAGG	4	2	6	1	80617	ribosomal protein S16
GCGACGAGGC	5	1	6	1	2017	ribosomal protein L38
TCAGACGCAG	2	4	6	1	250655	"prothymosin, alpha (gene sequence 28)"
TCAGTTTGGA	3	3	6	1	3873	"palmitoyl-protein thioesterase 1 (ceroid-lipofuscinosis, neuronal 1, infantile)"
TCGTGATTGT	2	4	6	1	125078	ornithine decarboxylase antizyme 1
AAGAGGCAAG	4	2	6	1		No Match

Tag Sequence	PMA	THP1	Total Tags	Redundant Matches	UniGene ID (Hs.)	Gene Description
AGGAAGGCGG	3	3	6	1		No Match
AGGAGGACTT	3	3	6	1		No Match
CAATAGAGAC	2	4	6	1		No Match
GAATGATCTG	3	3	6	1		No Match
GCCCGGGAAT	3	3	6	1		No Match
GTAAGCAAAA	0	6	6	1		No Match
GTGGTGCATA	2	4	6	1		No Match
TAATACTCAA	4	2	6	1		No Match
TATGTCAAGC	2	4	6	1		No Match
TCTGGACGCG	1	5	6	1		No Match
ATAGAGGCAA	2	4	6	1	173714	MORF-related gene X
AACGCTGCCA	1	5	6	1	301011	KIAA0876 protein
CAGAACCCAC	2	4	6	1	303627	"heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA-binding protein 1, 37kD)"
ATCCGAAAGA	1	5	6	1	322804	EST
ATTGTTTATG	3	3	6	2		Greater than one match
CCCGTGTGCT	4	2	6	2		Greater than one match
CCTAGCTGGA	5	1	6	2		Greater than one match
GAAAATATCC	0	6	6	2		Greater than one match
GAATTAACAT	3	2	5	1	79474	"tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide"
GGGGAAATCG	4	1	5	1	76293	"thymosin, beta 10"
GACTGAATCT	1	4	5	1	37936	suppressor of variegation 3-9 (Drosophila) homolog 1
TATCTGTCTA	3	2	5	1	145279	SET translocation (myeloid leukemia-associated)
GTGAAGGCAG	5	0	5	1	77039	ribosomal protein S3A
ATGGCTGGTA	2	3	5	1	182426	ribosomal protein S2
GTGGGCGTGT	4	1	5	1	133230	ribosomal protein S15
AAGACAGTGG	3	2	5	1	5566	ribosomal protein L37a
GAACACATCC	3	2	5	1	252723	ribosomal protein L19
CCCGTCCGGA	2	3	5	1	180842	ribosomal protein L13
CGCTGGTTCC	2	3	5	1	179943	ribosomal protein L11
ATGTGGTGTG	1	4	5	1	180909	peroxiredoxin 1
AACGAGGAAT	4	1	5	1		No Match
ACGCTGAATA	2	3	5	1		No Match
AGAGTTCAGA	2	3	5	1		No Match
AGCAATTCAA	4	1	5	1		No Match
AGCAGTCCCC	3	2	5	1		No Match
ATTCAAGACA	2	3	5	1		No Match
CAGGACTCCG	2	3	5	1		No Match
CTCTGACTTA	2	3	5	1		No Match
CTTAAGGATC	2	3	5	1		No Match
GAACAATGGA	3	2	5	1		No Match
GCGAAGCTCA	1	4	5	1		No Match
GGTAAGTGTG	4	1	5	1		No Match
TGGCTCGGTC	1	4	5	1		No Match
TGTACTCAAT	3	2	5	1		No Match
TGTAGTGTAA	2	3	5	1		No Match
TGTTCTATGG	2	3	5	1		No Match
TTATCAAGGG	3	2	5	1		No Match
TTGGCTGCCC	1	4	5	1		No Match
TTGGGCCAGA	3	2	5	1		No Match
TTGTGCAAAA	1	4	5	1		No Match

Tag Sequence	PMA	THP1	Total Tags	Redundant Matches	UniGene ID (Hs.)	Gene Description
GCGGTTGTGG	4	1	5	1	79356	Lysosomal-associated multispanning membrane protein-5
CTCTTCGAGA	4	1	5	1	76686	glutathione peroxidase 1
CCTACTAACC	4	1	5	1	128873	"ESTs, Highly similar to ALFA_HUMAN FRUCTOSE-BISPHOSPHATE ALDOLASE A [H.sapiens]"
AAGGACATCA	0	5	5	1	98110	ESTs
CCAGTCCTGG	1	4	5	1	221166	ESTs
GAGGTCACTG	2	3	5	1	136444	EST
GAAATATATG	1	4	5	1	429	"ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9) isoform 3"
AATTTCAAAA	3	2	5	2		Greater than one match
GAAATGTAAG	2	3	5	2		Greater than one match
GGATTTGGCT	2	3	5	2		Greater than one match
GGGCTGGGGT	2	3	5	2		Greater than one match
GTTCGTGCCA	3	2	5	2		Greater than one match
TCACCCACAC	4	1	5	2		Greater than one match
CAAAAATAAA	3	2	5	3		Greater than one match
GTGAAACCCC	5	0	5	127		Greater than one match

APPENDIX 4

NHMC 2° TRANSCRIPTOME

Descending Tag Frequency

Tag Sequence	Total LG	Total HG	Total	Induction	UniGene ID (Hs.)	Gene Description
Total Tags	23001	23128	46129			
ACAGGCTACG	282	280	562	0.99	75777	transgelin
ATGTGAAGAG	226	198	424	0.88	111779	"secreted protein, acidic, cysteine-rich (osteonectin)"
CACAAACGGT	139	127	266	0.91	195453	ribosomal protein S27 (metallopanstimulin 1)
TTTGCACCTT	131	129	260	0.98	75511	connective tissue growth factor
ATCTTGTTAC	120	135	255	1.13	287820	fibronectin 1
GACCGCAGGA	129	119	248	0.92	119129	"collagen, type IV, alpha 1"
GGAGTGTGCT	110	122	232	1.11	9615	"myosin regulatory light chain 2, smooth muscle isoform"
CATATCATTA	117	115	232	0.98	119206	insulin-like growth factor binding protein 7
AAGACAGTGG	112	87	199	0.78	296290	ribosomal protein L37a
TTGGTGAAGG	97	95	192	0.98	75968	"thymosin, beta 4, X chromosome"
CCCATCGTCC	96	86	182	0.9	mito	Tag matches mitochondrial sequence
GCCCCCAATA	88	93	181	1.06	227751	"lectin, galactoside-binding, soluble, 1 (galectin 1)"
TACCATCAAT	80	77	157	0.96	169476	glyceraldehyde-3-phosphate dehydrogenase
TGCCTCTGCG	79	72	151	0.91	75564	CD151 antigen
AGGCTACGGA	77	72	149	0.94	119122	ribosomal protein L13a
AGCACCTCCA	79	64	143	0.81	75309	eukaryotic translation elongation factor 2
GACCAGGCCC	73	65	138	0.89	300772	tropomyosin 2 (beta)
ATTCTCCAGT	82	54	136	0.66	234518	ribosomal protein L23
CCGTGACTCT	82	49	131	0.6	296267	follistatin-like 1
GCATAATAGG	68	57	125	0.84	184108	ribosomal protein L21
CGCCGCCGGC	60	63	123	1.05	182825	ribosomal protein L35
GTTGTGGTTA	56	59	115	1.05	75415	beta-2-microglobulin
CTGGGTTAAT	57	57	114	1	298262	ribosomal protein S19
TCAGATCTTT	60	52	112	0.87	108124	"ribosomal protein S4, X-linked"
ACATCATCGA	56	49	105	0.88	182979	ribosomal protein L12
CCCAAGCTAG	61	39	100	0.64	76067	heat shock 27kD protein 1
AAGGTGGAGG	57	41	98	0.72	163593	ribosomal protein L18a
GACGACACGA	47	50	97	1.06	153177	ribosomal protein S28
TAAGGAGCTG	54	43	97	0.8	299465	ribosomal protein S26
CCCGTCCGGA	58	39	97	0.67	180842	ribosomal protein L13
TGGTGTTGAG	58	33	91	0.57	275865	ribosomal protein S18
GTTTATGGAT	43	45	88	1.05	279009	matrix Gla protein
GGATTTGGCC	42	45	87	1.07	351937	"ribosomal protein, large P2"
TTCATACACC	35	48	83	1.37	mito	Tag matches mitochondrial sequence
GAGGGAGTTT	53	30	83	0.57	76064	ribosomal protein L27a
TTGGGGTTTC	44	38	82	0.86	62954	"ferritin, heavy polypeptide 1"
TGCATCTGGT	44	35	79	0.8	75410	"heat shock 70kD protein 5

Tag Sequence	Total LG	Total HG	Total	Induction	UniGene ID (Hs.)	Gene Description
						(glucose-regulated protein, 78kD)"
ACAGATTTGA	42	35	77	0.83	41271	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 1913076
GTTCGTGCCA	42	35	77	0.83	287361	ribosomal protein L35a
TTACCATATC	43	34	77	0.79	300141	ribosomal protein L39
TTGTTGTTGA	37	36	73	0.97	182278	"calmodulin 2 (phosphorylase kinase, delta)"
GGGGAAATCG	37	36	73	0.97	76293	"thymosin, beta 10"
CAATAAATGT	41	31	72	0.76	337445	ribosomal protein L37
GGCTGTACCC	41	30	71	0.73	108080	cysteine and glycine-rich protein 1
AGCCCTACAA	31	36	67	1.16	mito	Tag matches mitochondrial sequence
GATGAGGAGA	35	32	67	0.91	179573	"collagen, type I, alpha 2"
ACTTTTTCAA	37	30	67	0.81		
TCCCCGTAAT	40	27	67	0.68		
TAATAAAGGT	28	38	66	1.36	151604	ribosomal protein S8
AGGGCTTCCA	39	27	66	0.69	29797	ribosomal protein L10
GCCTGTATGA	34	31	65	0.91	180450	ribosomal protein S24
CGCTGGTTCC	31	31	62	1	179943	ribosomal protein L11
GGACCACTGA	30	30	60	1	119598	ribosomal protein L3
CACCTAATTG	28	31	59	1.11	mito	Tag matches mitochondrial sequence
CGAGGGGCCA	25	33	58	1.32	182485	"actinin, alpha 4"
ACTGAGGAAA	30	28	58	0.93	77326	insulin-like growth factor binding protein 3
AGGAAAGCTG	24	33	57	1.38	343443	ribosomal protein L36
CTGTTGGTGA	30	27	57	0.9	3463	ribosomal protein S23
ACTTGGAGTC	25	31	56	1.24	296842	"Homo sapiens cDNA: FLJ23324 fis, clone HEP12482, highly similar to HUMMYOHCB Human nonmuscle myosin heavy chain-B (MYH10) mRNA"
TGTGCTAAAT	25	31	56	1.24	250895	ribosomal protein L34
GAACACATCC	22	33	55	1.5	252723	ribosomal protein L19
AATCCTGTGG	31	23	54	0.74	178551	ribosomal protein L8
GGCTGGGGGC	36	18	54	0.5	75721	profilin 1
CCGTCCAAGG	22	31	53	1.41	80617	ribosomal protein S16
GCCCAAGGAC	26	27	53	1.04	195464	"filamin A, alpha (actin-binding protein-280)"
TTGTAATCGT	21	30	51	1.43	125078	ornithine decarboxylase antizyme 1
CTGCTATACG	21	30	51	1.43	180946	ribosomal protein L5
GAAGCAGGAC	23	28	51	1.22	180370	cofilin 1 (non-muscle)
ACCTGTATCC	27	24	51	0.89	182241	interferon induced transmembrane protein 3 (1-8U)
GGCAAGAAGA	29	22	51	0.76	111611	ribosomal protein L27
GCCGTGTCCG	32	19	51	0.59	350166	ribosomal protein S6
CTAAGACTTC	17	33	50	1.94		
TGCAATATGC	30	19	49	0.63	750	fibrillin 1 (Marfan syndrome)
TTGGAGATCT	30	18	48	0.6	50098	"NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4 (9kD, MLRQ)"

Tag Sequence	Total LG	Total HG	Total	Induction	UniGene ID (Hs.)	Gene Description
TCCCCGACAT	31	17	48	0.55		
TTAGTGTCGT	23	24	47	1.04	111779	"secreted protein, acidic, cysteine-rich (osteonectin)"
TGCACGTTTT	17	28	45	1.65	169793	ribosomal protein L32
TAAAAATGTT	19	25	44	1.32	82085	"serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1"
CCCCAGTTGC	17	26	43	1.53	74451	"calpain, small subunit 1"
AATAGGTCCA	18	25	43	1.39	113029	ribosomal protein S25
CCCATCCGAA	22	21	43	0.95	91379	ribosomal protein L26
ATCGTGGAGG	26	17	43	0.65	727	"inhibin, beta A (activin A, activin AB alpha polypeptide)"
TGGGCAAAGC	15	27	42	1.8	2186	eukaryotic translation elongation factor 1 gamma
CCCCCTGGAT	24	18	42	0.75	275243	S100 calcium-binding protein A6 (calcyclin)
GAGCCTGGAT	27	15	42	0.56	9004	chondroitin sulfate proteoglycan 4 (melanoma-associated)
ACTGGGTCTA	21	20	41	0.95	275163	"non-metastatic cells 2, protein (NM23B) expressed in"
GTGTTAACCA	17	23	40	1.35	74267	ribosomal protein L15
TAATGACAAT	26	14	40	0.54	239069	four and a half LIM domains 1
CTCAGACAGT	17	22	39	1.29	108957	40S ribosomal protein S27 isoform
GCTTTTAAGG	19	20	39	1.05	8102	ribosomal protein S20
AGCCTTTGTT	21	18	39	0.86	9930	"serine (or cysteine) proteinase inhibitor, clade H (heat shock protein 47), member 2"
TCACCGGTCA	13	25	38	1.92	290070	"gelsolin (amyloidosis, Finnish type)"
GGTTGGCAGG	15	23	38	1.53	3745	milk fat globule-EGF factor 8 protein
TCCCGTACAT	18	20	38	1.11		
GTTCCCTGGC	20	18	38	0.9	177415	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived); ribosomal protein S30
GGCCCTGAGC	23	15	38	0.65	71618	polymerase (RNA) II (DNA directed) polypeptide L (7.6kD)
CTGGGCGTGT	27	11	38	0.41	351987	"ESTs, Moderately similar to I60307 beta-galactosidase, alpha peptide - Escherichia coli [E.coli]"
GATCAGGCCA	16	21	37	1.31	119571	"collagen, type III, alpha 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)"
CTTTGAACGA	19	18	37	0.95	75511	connective tissue growth factor
GGTGGCACTC	15	21	36	1.4	77273	"ras homolog gene family, member A"

Tag Sequence	Total LG	Total HG	Total	Induction	UniGene ID (Hs.)	Gene Description
GTTTCATCTC	18	18	36	1	1940	"crystallin, alpha B"
ACAAGTACCC	20	16	36	0.8	142827	P311 protein
GTCTGGGGCT	23	13	36	0.57	75725	transgelin 2
AACGCGGCCA	19	16	35	0.84	73798	macrophage migration inhibitory factor (glycosylation-inhibiting factor)
TTTCTAGTTT	23	12	35	0.52	111894	lysosomal-associated protein transmembrane 4 alpha
TGGAAATGAC	11	23	34	2.09	172928	"collagen, type I, alpha 1"
GCCTTCCAAT	12	21	33	1.75	76053	"DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 5 (RNA helicase, 68kD)"
AATATGTGGG	13	20	33	1.54	351875	cytochrome c oxidase subunit VIc
TTTGCTCTCC	14	19	33	1.36	75350	vinculin
TGAGGGAATA	15	18	33	1.2	83848	triosephosphate isomerase 1
AGCAGATCAG	17	16	33	0.94	119301	"S100 calcium-binding protein A10 (annexin II ligand, calpactin I, light polypeptide (p11))"
GGGCGCTGTG	13	19	32	1.46	8372	ubiquinol-cytochrome c reductase (6.4kD) subunit
CTGACCTGTG	15	17	32	1.13	77961	"major histocompatibility complex, class I, B"
AAATGCCACA	17	15	32	0.88	65450	reticulon 4
ATCAAGGGTG	17	15	32	0.88	157850	ribosomal protein L9
TTATGTTTAA	13	18	31	1.38	79914	lumican
CTGAGAGCTG	14	17	31	1.21	78501	growth arrest-specific 6
GAAGAAATTA	18	13	31	0.72	325474	caldesmon 1
CAAACCATCC	18	13	31	0.72	65114	keratin 18
TCCCCGTACT	15	15	30	1		
TGTCATCACA	17	13	30	0.76	83354	lysyl oxidase-like 2
TCTTGTGCAT	15	14	29	0.93	2795	lactate dehydrogenase A
TCCAAATCGA	19	10	29	0.53	297753	vimentin
CCTCGGAAAA	9	19	28	2.11	2017	ribosomal protein L38
ACCAAAAACC	10	18	28	1.8	172928	"collagen, type I, alpha 1"
TACAAGAGGA	10	18	28	1.8	349961	ribosomal protein L6
AAAAGCTTGA	12	16	28	1.33	349933	"ESTs, Highly similar to A46546 leukocyte common antigen long splice form precursor [H.sapiens]"
TGGCCCCACC	13	15	28	1.15	198281	"pyruvate kinase, muscle"
GCCTGCTGGG	12	15	27	1.25	2706	glutathione peroxidase 4 (phospholipid hydroperoxidase)
CAAGCATCCC	18	9	27	0.5	mito	Tag matches mitochondrial sequence
ACAAACTTAG	13	13	26	1	177656	"calmodulin 1 (phosphorylase kinase, delta)"
CGACGAGGAG	14	12	26	0.86	9999	epithelial membrane protein 3
TAACTTGTGA	14	12	26	0.86	295726	"integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)"
TGAAGTTATA	15	11	26	0.73	287797	"integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)"

Tag Sequence	Total LG	Total HG	Total	Induction	UniGene ID (Hs.)	Gene Description
TACTTGGGAG	17	9	26	0.53	154103	LIM protein (similar to rat protein kinase C-binding enigma)
CCTCAGGATA	10	15	25	1.5		
ACTTACCTGC	11	14	25	1.27	174031	cytochrome c oxidase subunit VIb
TGGTACACGT	12	13	25	1.08	279574	cell death-regulatory protein GRIM19
TGTTCTGGAG	12	13	25	1.08	74471	"gap junction protein, alpha 1, 43kD (connexin 43)"
GGAGGGATCA	12	13	25	1.08	6196	integrin-linked kinase
ACTTATTATG	16	9	25	0.56	76152	decorin
GATAACTACA	9	15	24	1.67	119206	insulin-like growth factor binding protein 7
ATAGAGGCAA	10	14	24	1.4	173714	MORF-related gene X
TGGAGTGGAG	11	13	24	1.18	3764	guanylate kinase 1
GCACAAGAAG	14	10	24	0.71	289721	growth arrest-specific 5
TGTGAGCCCC	15	9	24	0.6	102948	enigma (LIM domain protein)
CGCCGCGGTG	9	14	23	1.56	4835	"eukaryotic translation initiation factor 3, subunit 8 (110kD)"
CCTGGAAGAG	10	13	23	1.3	75655	"procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), beta polypeptide (protein disulfide isomerase; thyroid hormone binding protein p55)"
TCAGACAAAA	12	11	23	0.92	66881	"dynein, cytoplasmic, intermediate polypeptide 2"
TCCGGCCGCG	12	11	23	0.92	171774	hypothetical protein
GAAGATGTGT	13	10	23	0.77	112318	6.2 kd protein
CCGGGTGATG	13	10	23	0.77	279910	"ATX1 (antioxidant protein 1, yeast) homolog 1"
GTGCTGGAGA	14	9	23	0.64	53125	small nuclear ribonucleoprotein D2 polypeptide (16.5kD)
CCCAGAGACC	19	4	23	0.21	21223	"calponin 1, basic, smooth muscle"
CAGCTCACTG	8	14	22	1.75	738	ribosomal protein L14
TTTAACGGCC	9	13	22	1.44	mito	Tag matches mitochondrial sequence
GTGCGCTAGG	11	11	22	1	9408	"Homo sapiens cDNA FLJ31238 fis, clone KIDNE2004864"
CAACTTAGTT	11	11	22	1	180224	myosin regulatory light chain
CTGCCAAGTT	12	10	22	0.83	75873	zyxin
AAGAACCTGT	14	8	22	0.57	75617	"collagen, type IV, alpha 2"
TAATATTTTT	6	15	21	2.5	182485	"actinin, alpha 4"
CCCTTAGCTT	7	14	21	2	180224	myosin regulatory light chain
ATGTCTTTTC	9	12	21	1.33	1516	insulin-like growth factor-binding protein 4
ACAACTCAAT	11	10	21	0.91	75922	brain protein I3
TCTGCCTATG	12	9	21	0.75	90291	"laminin, beta 2 (laminin S)"
GCGACCGTCA	13	8	21	0.62	273415	"aldolase A, fructose-bisphosphate"
AAAAAGCAGA	13	8	21	0.62	75428	"superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))"

Tag Sequence	Total LG	Total HG	Total	Induction	UniGene ID (Hs.)	Gene Description
AAAGTCTAGA	14	7	21	0.5	82932	cyclin D1 (PRAD1: parathyroid adenomatosis 1)
GTTAACGTCC	16	5	21	0.31	178391	ribosomal protein L44
GGCTGGTCTG	18	3	21	0.17	337986	hypothetical protein MGC4677
TGGGAAGTGG	8	12	20	1.5	112844	maternally expressed 3
TCTCTACCCA	11	9	20	0.82	279518	amyloid beta (A4) precursor-like protein 2
AAAACATTCT	11	9	20	0.82	mito	Tag matches mitochondrial sequence
TGGGAGGCTT	13	7	20	0.54	128151	ESTs
GCTTGGATCT	5	14	19	2.8		
AGAAAGATGT	6	13	19	2.17	78225	annexin A1
CCCCGCCAAG	6	13	19	2.17	169718	calponin 2
TTTTGGGGGC	6	13	19	2.17	46736	hypothetical protein FLJ23476
GCTTACCTTT	7	12	19	1.71	7753	calumenin
GACTGTGCCA	8	11	19	1.38	5120	"dynein, cytoplasmic, light polypeptide"
TTTTATGGAA	8	11	19	1.38	77269	"guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2"
TGATAATTCA	8	11	19	1.38	171625	hypothetical protein MGC14697
GACCTATCTC	8	11	19	1.38	194431	palladin
CAGGCCCCAC	8	11	19	1.38	256290	S100 calcium-binding protein A11 (calgizzarin)
ACGTTCTCTT	8	11	19	1.38		
GTTCAAAGAC	9	10	19	1.11	75260	mitogen inducible 2
AGGTCCTAGC	10	9	19	0.9	226795	glutathione S-transferase pi
GGAAATGTCA	10	9	19	0.9	111301	"matrix metalloproteinase 2 (gelatinase A, 72kD gelatinase, 72kD type IV collagenase)"
GAAATGATGA	11	8	19	0.73	288856	prefoldin 5
AACTAATACT	12	7	19	0.58	295362	DR1-associated protein 1 (negative cofactor 2 alpha)
TTACGAGGAA	12	7	19	0.58	227949	SEC13 (S. cerevisiae)-like 1
GAAACCGAGG	6	12	18	2	279813	hypothetical protein
TTCCGGTTCC	7	11	18	1.57	172609	nucleobindin 1
TACCTCTCTA	8	10	18	1.25	296842	"Homo sapiens cDNA: FLJ23324 fis, clone HEP12482, highly similar to HUMMYOHCB Human nonmuscle myosin heavy chain-B (MYH10) mRNA"
GTACTGTAGC	8	10	18	1.25	265829	"integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)"
AAACTTTGCC	9	9	18	1	194431	palladin
TGTTCATCAT	9	9	18	1	65450	reticulon 4
TAGTTGAAGT	9	9	18	1	131255	ubiquinol-cytochrome c reductase binding protein
TTTTTGTACA	11	7	18	0.64	78040	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 1

APPENDIX 5

FULL NHMC DIFFERENTIAL LIST

Alphabetical by Gene Description

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
GCGACGGCCG		4	1	-4	0.08	112318	6.2 kd protein
CACTCCAGCC		4	0	-4	0.15	108258	actin binding protein; macrophin
ACAAACTGTG		1	4	4	0.08	90370	actin related protein 2/3 complex, subunit 1A (41 kD)
TCTTTACTTG		4	0	-4	0.15	6895	actin related protein 2/3 complex, subunit 3 (21 kD)
AACATCAAAC		1	5	5	0.11	82425	actin related protein 2/3 complex, subunit 5 (16 kD)
AAGATCAAGG		0	5	5	0.19	1288	actin, alpha 1, skeletal muscle
ATTTTGTGTC		1	5	5	0.11	75056	adaptor-related protein complex 3, delta 1 subunit
AAAACATTAT		6	1	-6	0.14	80917	adaptor-related protein complex 3, sigma 1 subunit
CTCATCAGCT		0	6	6	0.23	104125	adenylyl cyclase-associated protein
TCTTTGCTCT		4	0	-4	0.15	44077	alpha-parvin
CTGGCGCCGA		1	6	6	0.14	183180	anaphase promoting complex subunit 11
AAAATAAAGA		1	5	5	0.11	73722	APEX nuclease (multifunctional DNA repair enzyme)
TGGACACAAG		4	1	-4	0.08	180832	arginyl-tRNA synthetase
AGGTGCGGGG		4	1	-4	0.08	165439	arsA (bacterial) arsenite transporter
TCCCTGTTAA		0	4	4	0.15	75415	beta-2-microglobulin
TCCGTGGTTG		1	5	5	0.11	79516	brain acid-soluble protein 1
CACCACGGGC		1	5	5	0.11	273219	breast cancer anti-estrogen resistance 1
ACGAATATCA		4	0	-4	0.15	63984	cadherin 13, H-cadherin (heart)
CACACACACA	C	4	0	-4	0.15	63984	cadherin 13, H-cadherin (heart)
ATCCGTGCCC		4	1	-4	0.08	141011	calmodulin 3 (phosphorylase kinase, delta)
CCCAGAGACC		15	4	-3.75	0.2	21223	calponin 1, basic, smooth muscle
CCGTGCTCAT		4	1	-4	0.08	9857	carbonyl reductase
GCAGCTCAGG		5	0	-5	0.19	79572	cathepsin D (lysosomal aspartyl protease)
TTGGTGGAGG		4	0	-4	0.15	76294	CD63 antigen (melanoma 1 antigen)
TCTGTTCTGG		4	1	-4	0.08	76932	cell division cycle 34
TCTCAATTCT	T	1	8	8	0.21	146409	cell division cycle 42

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
							(GTP-binding protein, 25kD)
TTAATAAAAG	T	2	8	4	0.13	152738	CGI-113 protein
TAACTTAAGC		1	5	5	0.11	184542	CGI-127 protein
TGTGTGCCAC	T	4	0	-4	0.15	72925	chromosome 11 open reading frame 13
AGGGTTGGAA		4	0	-4	0.15	15106	chromosome 14 open reading frame 1
GAGAATCTGC	T	4	0	-4	0.15	179260	chromosome 14 ORF 4
AGTTCCACCA		1	4	4	0.08	182626	chromosome 22 open reading frame 5
TTTCCTTCCT		7	1	-7	0.17	104143	clathrin, light polypeptide (Lca)
TACCAAGACC		4	0	-4	0.15	3059	coatomer protein complex, subunit beta
CCATTTTCTG	G	0	5	5	0.19	83164	collagen, type XV, alpha 1
AAATGGATAC		11	3	-3.67	0.15	75511	connective tissue growth factor
CACCCCTGAT		0	4	4	0.15	173724	creatine kinase, brain
GAGAATCTGC	T	4	0	-4	0.15	23960	cyclin B1
ATGAGCTGAC		1	7	7	0.17	695	cystatin B (stefin B)
GACCAGAAAA		4	0	-4	0.15	180714	cytochrome c oxidase subunit VIa polypeptide 1
TTAAACTCTA		4	1	-4	0.08	226213	cytochrome P450, 51
CCACTCCTCC		1	4	4	0.08	82890	defender against cell death 1
AAAGTTCGTA		1	10	10	0.3	82306	destrin (actin depolymerizing factor)
GTGACTGCCA	C	5	1	-5	0.11	84183	diptheria toxin resistance protein required for diphthamide biosynthesis
TTGTCAATGG	G	7	2	-3.5	0.1	239370	DKFZP727I051 protein
AGACCAAAGT		4	1	-4	0.08	82646	DnaJ (Hsp40) homolog, subfamily B, member 1
AATGCTGGCA		5	0	-5	0.19	181195	DnaJ (Hsp40) homolog, subfamily B, member 6
AAGGAGAAGG		1	5	5	0.11	34789	dolichyl-diphosphooligosaccharide-protein glycosyltransferase
AACTATACAA		1	4	4	0.08	15432	downregulated in ovarian cancer 1
ATTTCTCATT		0	5	5	0.19	36794	D-type cyclin-interacting protein 1
ACGCCCTGCT		7	1	-7	0.17	898	dystrophia myotonica-protein kinase
CAGTTGGTTG		4	0	-4	0.15	155218	E1B-55kDa-associated protein 5
AAACATTGGG		6	1	-6	0.14	8203	endomembrane protein emp70 precursor isolog

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
AAAAAACCCA		4	1	-4	0.08	111680	endosulfine alpha
CTGTGCTCGG		1	4	4	0.08	76394	enoyl Coenzyme A hydratase, short chain, 1, mitochondrial
GACAGGGCTG	G	4	0	-4	0.15	307452	EST
GGCCAAAGGC	C	6	0	-6	0.23	213701	EST
CAAATCAAGT		0	5	5	0.19	296234	ESTs
GATTAAGTGA		1	5	5	0.11	95835	ESTs
GTGCTTGAAT	G	3	13	4.33	0.21	313280	ESTs
AAAAATGGTG		1	4	4	0.08	204930	ESTs
ACTTTTCAAA		0	4	4	0.15	44609	ESTs
AACAGAAGCA		7	2	-3.5	0.1	292815	ESTs
GACAGGGCTG	G	4	0	-4	0.15	186669	ESTs
CCCTCTTTGG		4	0	-4	0.15	181174	ESTs
TGTGTGCCAC	T	4	0	-4	0.15	153136	ESTs
ATACATACTG		4	1	-4	0.08	74313	ESTs
ATGTGGGTCT		4	0	-4	0.15	42392	ESTs
CCTGTCTGCA		4	1	-4	0.08	25338	ESTs
ATTTGTATCT						292905	ESTs
ACGGCTCCGA		1	5	5	0.11	48563	ESTs, Moderately similar to ALU2
ATTAAACTTG	G	0	5	5	0.19	323908	ESTs, Weakly similar to Z192_HUMAN ZINC FINGER PROTEIN 192
TGGTTTGCAC	A	1	4	4	0.08	3593	ESTs-Similar to lysozyme
CCATTTTCTG	G	0	5	5	0.19	198899	eukaryotic translation initiation factor 3, subunit 10
ACACAGTTTT		1	5	5	0.11	166994	FAT tumor suppressor (Drosophila) homolog
TATCTGATCT		0	4	4	0.15	166994	FAT tumor suppressor (Drosophila) homolog
ACTGAGGTGC		4	1	-4	0.08	7768	fibroblast growth factor (acidic) intracellular binding protein
GCTTGGATCT		4	14	3.5	0.17	250723	FK506 binding protein 12-rapamycin associated protein 1
GAATAAATGT		1	5	5	0.11	302749	FK506-binding protein 9 (63 kD)
CCTGTTCTCC		5	1	-5	0.11	109798	G8 protein
AAGAAGACTT		1	4	4	0.08	7719	GABA(A) receptor-associated protein
CCTTTCCTTT		1	4	4	0.08	74576	GDP dissociation inhibitor 1
GGCCAGCAAT	T	1	4	4	0.08	64639	glioma pathogenesis-related protein
GAGTAAAAAA	T	1	4	4	0.08	180532	glucose phosphate isomerase
ACAGGCTCGG		3	10	3.33	0.12	597	glutamic-oxaloacetic transaminase 1, soluble
AACTAAAAAA	A	1	7	7	0.17	55921	glutamyl-prolyl-tRNA synthetase
CTCTTCGAGA		8	2	-4	0.13	76686	glutathione peroxidase 1

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
AGGGAGGGGC		4	1	-4	0.08	172153	glutathione peroxidase 3 (plasma)
ACCATCAATA		1	8	8	0.21	169476	glyceraldehyde-3-phosphate dehydrogenase
TAGACCCCTT		2	7	3.5	0.1	169476	glyceraldehyde-3-phosphate dehydrogenase
CCCATCATCC		5	0	-5	0.19	306122	glycoprotein, synaptic 2
CCAACCGTGC		7	2	-3.5	0.1	75207	glyoxalase I
GTGTCCTCCT		0	4	4	0.15	78979	Golgi apparatus protein 1
AGCCTTCCTA	G	4	0	-4	0.15	78979	Golgi apparatus protein 1
TTAATAGTGG		0	4	4	0.15	18271	Golgi protein
ACCTGCTGGT		5	1	-5	0.11	5807	GTPase Rab14
TCCCCACATC		0	5	5	0.19	51147	guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1
GAAACCAACT		5	0	-5	0.19	296261	guanine nucleotide binding protein (G protein), q polypeptide
ACACTTCTTT		1	4	4	0.08	83381	guanine nucleotide binding protein 11
CCCACACTAC		6	1	-6	0.14	91299	guanine nucleotide binding protein, beta polypeptide 2
TTTCCTTCCT		7	1	-7	0.17	285688	H.sapiens clathrin light chain
GAAATTTAAA		1	4	4	0.08	274472	high-mobility group protein 1
ATTTGTCCCA		4	0	-4	0.15	139800	high-mobility group protein isoforms I and Y
TTTACAAGTT		0	4	4	0.15	104640	HIV-1 inducer of short transcripts binding protein
GCTTCCATCT	G0/2, T0/4	1	4	4	0.08	55296	HLA-B associated transcript-1
ATTAAAGTGC		4	1	-4	0.08	63243	Homo sapiens cDNA FLJ10041 fis, clone HEMBA1001022
GGCTGGTCTG		17	3	-5.67	0.37	50724	Homo sapiens cDNA FLJ10934
TCTCTACTAA	A	1	5	5	0.11	314347	Homo sapiens cDNA FLJ11507
ATTTGTATCT	A	1	4	4	0.08	281434	Homo sapiens cDNA FLJ14028 fis
TCACCCACAC	C	6	23	3.83	0.28	322680	Homo sapiens cDNA: FLJ21547
CCAGCCTGGG		5	1	-5	0.11	306862	Homo sapiens cDNA: FLJ23014 fis
GCGGGGTACC		1	6	6	0.14	322466	Homo sapiens cDNA: FLJ23491 fis
GGGGATGGGG	T	6	1	-6	0.14	99093	Homo sapiens chromosome 19,

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
							cosmid R28379
TCCTGTAAAG		1	5	5	0.11	74034	Homo sapiens clone 24651 mRNA sequence
TACATCCGAA		5	1	-5	0.11	21321	Homo sapiens clone FLB9213 PRO2474 mRNA
GTGCTCTGTA	C	1	4	4	0.08	247993	Homo sapiens HLA class III region containing NOTCH4,PBX2 (HPBX) ,RAGE
GGGGATGGGG	T	6	1	-6	0.14	284278	Homo sapiens HSPC080 mRNA, partial cds
ACAAAGTAGG		4	1	-4	0.08	182183	Homo sapiens mRNA for caldesmon, 3' UTR
CTTCTTCTGT		0	4	4	0.15	102367	Homo sapiens mRNA for HMG-box transcription factor TCF-3
ATTTCACATT		1	5	5	0.11	7378	Homo sapiens mRNA; cDNA DKFZp434G227
GTGTTCCCAT		0	4	4	0.15	267120	Homo sapiens mRNA; cDNA DKFZp434O1427
AACACAATCA		4	1	-4	0.08	321403	Homo sapiens mRNA; cDNA DKFZp564O2363
AAAGGAATAA	T	1	4	4	0.08	172089	Homo sapiens mRNA; cDNA DKFZp586I2022
GTGCTCTGTA						322456	Homo sapiens mRNA; cDNA DKFZp761D0211
TTAATAAAAG	T	2	8	4	0.13	285902	Homo sapiens T-cell activation protein (PGR1) gene
CAGATGCAAA		2	7	3.5	0.1	94695	Homo sapiens TLH29 protein precursor (TLH29) mRNA
TGAATGGCCT		1	6	6	0.14	20597	host cell factor homolog
GAGGGCCGTG		5	1	-5	0.11	25635	HSPC003 protein
ATTCAGCACC		4	1	-4	0.08	11125	HSPC033 protein
GGAGATGGAG		0	5	5	0.19	188757	Human DNA sequence from clone 108K11 on chromosome 6p21
TAGGGCAATC		4	1	-4	0.08	113293	Human DNA sequence from clone 281H8 on chromosome 6q25.1-25.3.
GCTGACTCAG	G	4	0	-4	0.15	105607	Human DNA sequence from clone RP4-794I6 on chromosome 20
CGGCTCAAGT	C	0	4	4	0.15	279868	hypothetical protein
TCTGTGCTCA		1	4	4	0.08	22129	hypothetical protein
AAAACTGCAC		4	0	-4	0.15	8084	hypothetical protein

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
							dJ465N24.2.1
ACTTTTCAAA		0	4	4	0.15	33032	hypothetical protein DKFZp434N185
GGGCCCAGGA		1	4	4	0.08	118983	hypothetical protein FLJ12150
TTCCTCCACG	C	1	4	4	0.08	16603	hypothetical protein FLJ13163
CTGGGGGTCT		0	4	4	0.15	15356	hypothetical protein FLJ20254
TACACCAAGA		0	4	4	0.15	6449	hypothetical protein FLJ20542
AGTGTTTGTA		0	4	4	0.15	5420	hypothetical protein FLJ20695
GAGGCGATCA		1	4	4	0.08	30783	hypothetical protein FLJ20850
TCTTCGTCCT		4	0	-4	0.15	14891	hypothetical protein FLJ21047
GAAGTGGAAG	C	1	4	4	0.08	9061	hypothetical protein MGC2477
GGCACAGTAA		1	5	5	0.11	11270	hypothetical protein MGC2491
GTGTGTGGTG	C	1	4	4	0.08	151032	hypothetical protein MGC2683
TTCATTAAAA		1	4	4	0.08	287797	integrin, beta 1
CGCCGACGAT		9	2	-4.5	0.16	265827	interferon, alpha-inducible protein
TGACACCCAC		5	0	-5	0.19	76038	isopentenyl-diphosphate delta isomerase
GAACAGTGTG		0	4	4	0.15	91143	jagged 1 (Alagille syndrome)
GAAAGGTCTG		1	4	4	0.08	118778	KDEL (Lys-Asp-Glu-Leu) ER protein retention rc2
GGGCCCCGCA		5	1	-5	0.11	75353	KIAA0123 protein
TATTAACTCT		4	1	-4	0.08	45180	KIAA0337 gene product
AGTTGTCCCG		1	5	5	0.11	297641	KIAA0462 protein
GGGGCTGGAG		4	1	-4	0.08	301685	KIAA0620 protein
TGATCCATCC		4	1	-4	0.08	178121	KIAA0626 gene product
CAAATAAATG	T	4	0	-4	0.15	6654	KIAA0657 protein
TGTGCTTTTT	T	4	0	-4	0.15	22039	KIAA0758 protein
CTGGGTTGTG		0	4	4	0.15	10669	KIAA1249 protein
TCTTCTGCCA		4	1	-4	0.08	86392	KIAA1402 protein
GTGGAATAAA		0	4	4	0.15	83337	latent transforming growth factor beta binding protein 2
CCCTCTCCCT		4	0	-4	0.15	85087	latent transforming growth factor beta binding protein 4
AGCCTTCCTA	G	4	0	-4	0.15	166318	lipin 2
TATCACTCTG		6	1	-6	0.14	278362	male-enhanced antigen
TGTCGCTGGG		7	2	-3.5	0.1	227152	mannan-binding lectin serine protease 1
GGGAGGGGTG	G	9	2	-4.5	0.16	2399	matrix metalloproteinase 14 (membrane-inserted)
AGGACAGAAG	G	1	4	4	0.08	198265	matrix

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
							metalloproteinase 25
AAGATAATGC		5	0	-5	0.19	102696	MCT-1 protein
CTTTTCAAGA	A	4	0	-4	0.15	83532	membrane cofactor protein (CD46, trophoblast-lymphocyte cross-reactive antigen)
ATTACACCAC		4	1	-4	0.08	107014	membrane interacting protein of RGS16
GGTGGATGTG	C	4	1	-4	0.08	178728	methyl-CpG binding domain protein 3
GGGAATAAAC		4	1	-4	0.08	3828	mevalonate (diphospho) decarboxylase
AAATAAAGAA		4	1	-4	0.08	790	microsomal glutathione S-transferase 1
AAGTATGTGA		4	0	-4	0.15	75260	mitogen inducible 2
GCCCCCCACT		4	1	-4	0.08	75074	mitogen-activated protein kinase-activated protein kinase 2
GTGTGTGGTG						284203	myogenic factor 3
ATTTGAGAGT		2	7	3.5	0.1	146550	myosin, heavy polypeptide 9, non-muscle
TGCTACGAAA		6	1	-6	0.14	146550	myosin, heavy polypeptide 9, non-muscle
GTGCTTGAAT	G	3	13	4.33	0.21	77385	myosin, light polypeptide 6, alkali, smooth muscle and non-muscle
CAATGTGCTG		4	0	-4	0.15	141727	myotubularin related protein 4
AGCTGATCAG		4	0	-4	0.15	78223	N-acylaminoacyl-peptide hydrolase
CAATGTGTTA		1	8	8	0.21	74823	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1 (7.5kD, MWFE)
GGGAGCTGCG		1	7	7	0.17	211914	NADH dehydrogenase (ubiquinone) Fe-S protein 7
AAGTCATTCA	G	1	5	5	0.11	274416	NADH dehydrogenase 1 alpha subcomplex, 6
AAATACTGCC		5	0	-5	0.19	32916	nascent-polypeptide-associated complex alpha polypeptide
ATACAAGAGC		1	10	10	0.3	-	No Match
CTTCAGCTAA		0	6	6	0.23	-	No Match
TCCCTACATC		0	5	5	0.19	-	No Match
TCCTCGTACA		0	5	5	0.19	-	No Match
AGGACAGTGG		1	5	5	0.11	-	No Match
GCAAGCCATC		1	5	5	0.11	-	No Match
GACGACACGG		0	4	4	0.15	-	No Match
GTGAAAACCC		0	4	4	0.15	-	No Match
GTGCTGAACG		0	4	4	0.15	-	No Match
AGCTTATACT		1	4	4	0.08	-	No Match

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
AATATTATAT		4	0	-4	0.15	-	No Match
AGCACCTCAG		4	0	-4	0.15	-	No Match
GTTCTCCAGT		4	1	-4	0.08	-	No Match
CCCTGCACTC		7	1	-7	0.17	-	No Match
GTGCTGAAGG		9	1	-9	0.26	-	No Match
ATGACTCAAG		6	1	-6	0.14	239752	nuclear receptor subfamily 2, group F, member 6
TACAAAACCA		1	6	6	0.14	79110	nucleolin
GCACCTTATT		2	11	5.5	0.23	125078	ornithine decarboxylase antizyme 1
ACTACTAAGG		10	3	-3.33	0.12	2820	oxytocin receptor
AAGTCATTCA	G	1	5	5	0.11	9629	papillary renal cell carcinoma (translocation-associated)
GACTCACTTT		2	11	5.5	0.23	699	peptidylprolyl isomerase B (cyclophilin B)
GGCCAGCCCT	T	1	4	4	0.08	155455	phosphofructokinase, liver
TAACCCAACA		5	0	-5	0.19	1869	phosphoglucomutase 1
CTTATTTGTT		4	1	-4	0.08	4114	plastin 3 (T isoform)
GCGGGGTACC		1	6	6	0.14	227823	pM5 protein
CCTGCCAAAG		4	0	-4	0.15	75323	prohibitin
ACATACAACT		4	1	-4	0.08	61153	proteasome (prosome, macropain) 26S subunit, ATPase, 2
TTCACAAAGG		6	1	-6	0.14	76913	proteasome (prosome, macropain) subunit, alpha type, 5
ATCAGTGGCT	G1/T 6	1	7	7	0.17	89545	proteasome (prosome, macropain) subunit, beta type, 4
GTGCTGGACC	T	1	5	5	0.11	179774	proteasome activator subunit 2
AATGACTGAA		7	2	-3.5	0.1	93659	protein disulfide isomerase related protein
ATCCAGGGTC		4	0	-4	0.15	93659	protein disulfide isomerase related protein
TGTGCTTTTT	T	4	0	-4	0.15	77271	protein kinase, cAMP-dependent, catalytic, alpha
AAGATTTTAG		0	6	6	0.23	21537	protein phosphatase 1, catalytic subunit, beta isoform
CACACACACA	C	4	0	-4	0.15	127614	protein phosphatase 1, regulatory (inhibitor) subunit 3
AAAATGCTGG	T	0	3	3	0.11	57764	protein phosphatase 1A (formerly 2C), magnesium-dependent, alpha isoform
GTCTGACCCC		0	5	5	0.19	173902	protein phosphatase 2 (formerly 2A)
GTGGGGCTAG		0	4	4	0.15	75180	protein phosphatase 5,

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
							catalytic subunit
GGTGATGAGG		1	4	4	0.08	12107	putative breast adenocarcinoma marker (32kD)
CCACCGCACT		5	1	-5	0.11	115325	RAB7, member RAS oncogene family-like 1
CTGGATGCCG		5	1	-5	0.11	106061	RD RNA-binding protein
ACTTTCAGAT		3	11	3.67	0.15	227571	regulator of G-protein signalling 4
TAAAATGAAA		0	9	9	0.38	24950	regulator of G-protein signalling 5
ATAATAAAGC		4	0	-4	0.15	37682	retinoic acid receptor responder (tazarotene induced) 2
TATTTCACCG		4	1	-4	0.08	138860	Rho GTPase activating protein 1
GCAAGCCCCA		0	4	4	0.15	252574	ribosomal protein L10a
TCACCCACAC	C	6	23	3.83	0.28	234518	ribosomal protein L23
GCGACAGCTC		6	0	-6	0.23	184582	ribosomal protein L24
GTTAACGTCC		14	4	-3.5	0.17	178391	ribosomal protein L44
AACTAAAAAA	A	1	7	7	0.17	3297	ribosomal protein S27a
AACTAACAAA		3	10	3.33	0.12	3297	ribosomal protein S27a
TTTCTTAAAG		1	4	4	0.08	197114	RNA binding protein; AT-rich element binding factor
TTGAATTTGT		4	1	-4	0.08	80248	RNA-binding protein gene with multiple splicing
ATGTGAGGAG		0	5	5	0.19	111779	secreted protein, acidic, cysteine-rich (osteonectin)
TTGAATTCCC		5	1	-5	0.11	171921	sema domain, immunoglobulin domain (Ig)
ATGATGCGGT		7	1	-7	0.17	41072	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 6
GGTTATTTTG		2	7	3.5	0.1	82085	serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
ATGGCCATAG		4	0	-4	0.15	155206	serine/threonine kinase 25
GTGGATGGAC		4	0	-4	0.15	6418	seven transmembrane domain orphan receptor
TTTCAGAGAG		5	1	-5	0.11	75975	signal recognition particle 9kD
CCGGAAACAC		1	5	5	0.11	288013	similar to yeast BET3 (S. cerevisiae)
TAGGGCAATC		4	1	-4	0.08	180139	SMT3 (suppressor of mif two 3, yeast)

Tag Seq	11th Base	LG	HG	Induction	Beta (0.75,6,6)	UniGene ID (Hs.)	Gene Description
							homolog 2
GAGACTCCTG		4	0	-4	0.15	169902	solute carrier family 2 (facilitated glucose transporter), member 1
GGTGAGACAC		4	1	-4	0.08	164280	solute carrier family 25, member 6
GGCCAAAGGC						278569	sorting nexin 17
TAAAAGACAA		1	8	8	0.21	77196	spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
TTGATGTACA		8	2	-4	0.13	11482	splicing factor, arginine/serine-rich 11
TTCCTCCACG	C	1	4	4	0.08	183373	src homology 3 domain-containing protein HIP-55
GATTGAACCT		1	4	4	0.08	239926	sterol-C4-methyl oxidase-like
GACTGTTAAT		0	6	6	0.23	118684	stromal cell-derived factor 2
CGGCTCAAGT	C	0	4	4	0.15	250747	SUMO-1 activating enzyme subunit 1
TGGCCTAATA		4	1	-4	0.08	1501	syndecan 2 (heparan sulfate proteoglycan 1)
CCACCCCGAA		4	0	-4	0.15	74637	testis enhanced gene transcript (BAX inhibitor 1)
GCTGACTCAG	G	4	0	-4	0.15	87409	thrombospondin 1
AATGCAAGAT		1	4	4	0.08	171626	transcription elongation factor B (SIII)
TTATGTATCA		0	6	6	0.23	169300	transforming growth factor, beta 2
TACGTTGCAG		0	4	4	0.15	21756	translation factor sui1 homolog
CAGATAACAT		4	1	-4	0.08	75187	translocase of outer mitochondrial membrane 20 (yeast) homolog
TCTCTACTAA	A	1	5	5	0.11	250641	tropomyosin 4
CTGTACAGAC		2	8	4	0.13	251653	tubulin, beta, 2
TGTCTTTGCT		4	1	-4	0.08	9589	ubiquilin 1
GTAAAGAATA		0	4	4	0.15	80120	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 1 (GalNAc-T1)
CAGCCTCCCT		4	0	-4	0.15	75593	uroporphyrinogen III synthase (congenital erythropoietic porphyria)
CTGGGCCAGC		4	1	-4	0.08	74669	vesicle-associated membrane protein 5 (myobrevin)
CCTGAGCCCG		1	4	4	0.08	68571	VPS28 protein
AGCAGCGTGG		0	5	5	0.19	32117	xylosyltransferase II
TCCTCCCTAC		10	3	-3.33	0.12	70266	yeast Sec31p homolog

APPENDIX 6

TOP 200 GENES DETECTED IN GF200 ANALYSIS

Descending Relative Abundance

UniGene	Acc	Gene Description	Gene	LG	HG
Hs.140	N92646	"Human (hybridoma H210) anti-hepatitis A IgG variable region, constant region, complementarity-determining regions mRNA, complete cds"		140840.02	139906.33
Hs.23454	AA292995	"Dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)"	DCT	109017.29	123220.63
Hs.10761	R32802	"Homo sapiens secretory carrier membrane protein (SCAMP2) mRNA, complete cds"		100579.84	108898.29
Hs.8148	R78516	"ESTs, Weakly similar to C35C5.3 [C.elegans]"		85992.23	109749.13
Hs.105679	AA488072	ESTs		97576.66	97627.65
Hs.43704	AA046690	KINESIN HEAVY CHAIN		89245.28	92327.33
Hs.93231	W15465	ESTs		81729.84	92737.72
Hs.73527	R69566	ESTs		78608.4	91061.82
Hs.66576	AA281635	"Human MDA-7 (mda-7) mRNA, complete cds"		72013.8	92405.53
Hs.83262	AA443506	"Cell division cycle 42 (GTP-binding protein, 25kD)"	CDC42	71162.84	89656.99
Hs.82646	AA481758	DNAJ PROTEIN HOMOLOG 1		73669.79	86947.13
Hs.75690	AA019482	"Human mitochondrial creatine kinase (CKMT) gene, complete cds"		72777.27	81428.13
Hs.75721	AA521431	Desmin	DES	67233.21	82572.61
Hs.80021	N57731	ESTs		70618	75980.92
Hs.77550	AA459292	CDC28 protein kinase 1	CKS1	65775.6	80225.3
Hs.18203	AA411554	"Human GT mitochondrial solute carrier protein homologue mRNA, complete cds"		73755.21	58793.47
Hs.1726	W44701	"Interleukin 3 receptor, alpha (low affinity)"	IL3RA	64452.37	65826.84
Hs.24553	R32754	ESTs		64850.41	64683.77
Hs.23100	H93118	"ESTs, Weakly similar to similar to acyl-CoA dehydrogenases and epoxide hydrolases [C.elegans]"		55459.37	68054.59
Hs.71989	R67602	ESTs		54008.44	66651.93
Hs.94672	H94857	GCN5-like 1		57716.66	60766.84
Hs.73151	T81764	Cell division cycle 27	CDC27	58756.2	59398.06
Hs.64264	AA478436	"Human SWI/SNF complex 60 KDa subunit (BAF60b) mRNA, complete cds"		56182.13	61022.27
Hs.15219	AA196465	"Human sarcolipin (SLN) mRNA, complete cds"		48891.36	66995.56
Hs.91161	AA253430	"Human C-1 mRNA, complete cds"		50842.08	63920.38
Hs.26126	AA480851	"Homo sapiens putative OSP like protein mRNA, partial cds"		53362.12	54876.65
Hs.93097	R76499	ESTs		53111.63	54136.84
Hs.41874	H93842	ESTs		50412.42	54869.6
Hs.921	N62179	"Human methylmalonate semialdehyde dehydrogenase gene, complete cds"		48633.76	56258.47
Hs.119564	AA411554	"ESTs, Weakly similar to !!!! ALU SUBFAMILY SX WARNING ENTRY !!!! [H.sapiens]"		51234.48	52916.68
Hs.115756	AA458472	"Human MHC class II HLA-DR2-Dw12 mRNA DQw1-beta, complete cds"		51508.18	52586.09
RG.4	R70462			47393.37	56601.83
Hs.90283	AA490172	"Collagen, type I, alpha-2"	COL1A2	44795.5	59122.5
Hs.28426	R63530	ESTs		57043.33	45821.78
Hs.75184	AA434115	CARTILAGE GLYCOPROTEIN-39 PRECURSOR		45429.31	57270.89

UniGene	Acc	Gene Description	Gene	LG	HG
Hs.49093	N94234	ESTs		47909.96	54373.09
Hs.55921	AA599158	MULTIFUNCTIONAL AMINOACYL-TRNA SYNTHETASE		50376.54	51321.75
Hs.7107	R68464	ESTs		44652.13	55293.65
Hs.82535	N49856	SODIUM- AND CHLORIDE-DEPENDENT BETAINE TRANSPORTER		53281.7	45056.96
Hs.26275	W07300	"ESTs, Moderately similar to GAMMA-ADAPTIN [M.musculus]"		47485.54	48600.36
Hs.75984	R62612	Fibronectin 1	UBA52	45975.68	49052.8
Hs.24297	AA261796	Multiple endocrine neoplasia I	MEN1	44931.05	48883.45
Hs.100056	R62612	Fibronectin 1	FN1	44654.99	47366.23
Hs.3197	N29376	Myeloid cell nuclear differentiation antigen	MNDA	44290.5	46176.64
Hs.38768	AA459263	"Human Bcl-2 related (Bfl-1) mRNA, complete cds"		40813.97	49061.89
Hs.88611	AA279429	Endothelin converting enzyme 1	ECE1	41914.48	47905.14
Hs.2795	AA489611	Lactate dehydrogenase A	LDHA	43187.98	46140.77
Hs.77385	AA488346	"MYOSIN LIGHT CHAIN ALKALI, NON-MUSCLE ISOFORM"		41242.25	46842.3
Hs.82399	AA504526	H.sapiens LDLC mRNA		41000.17	45425.23
Hs.78054	AA456352	"Human mRNA for KIAA0224 gene, complete cds"		43114.22	43286.07
Hs.1242	AA464755	"Ankyrin 1, erythrocytic"	ANK1	41465.8	43508.42
Hs.75683	AA411440	Villin 2 (ezrin)	VIL2	38881.05	45145.16
Hs.19555	AA486332	Human clone 23867 mRNA sequence		40348.32	41917.11
Hs.30194	AA030013	"ESTs, Highly similar to PHOSPHATIDYLCHOLINE TRANSFER PROTEIN [Bos taurus]"		38088.33	43083.07
Hs.118973	AA280676	"ESTs, Weakly similar to X-linked mental retardation candidate gene [H.sapiens]"		42272.52	38795.28
Hs.78881	AA282642	MYOCYTE-SPECIFIC ENHANCER FACTOR 2		39330.6	40722.13
Hs.1162	AA280677	"Major histocompatibility complex, class II, DM beta"	HLA-DMB	39606.16	40415.81
Hs.41891	H93450	"ESTs, Weakly similar to ZINC FINGER PROTEIN 165 [H.sapiens]"		41994.98	36669.13
Hs.41817	H93906	ESTs		40493.94	38087.36
Hs.109052	T90621	"ESTs, Highly similar to 6.8 KD MITOCHONDRIAL PROTEOLIPID [Bos taurus]"		30765.18	45340.72
Hs.944	AA401111	Glucose phosphate isomerase	GPI	36438.63	38541.58
Hs.2175	AA443000	Colony stimulating factor 3 receptor (granulocyte)	CSF3R	31350.29	42455.95
Hs.105976	AA290737	Glutathione S-transferase M4	GSTM4	33352.9	39576.2
Hs.693	AA291995	"Cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kD"	CSTF2	35290.7	37145.03
Hs.78068	AA427724	"Homo sapiens carboxypeptidase Z precursor, mRNA, complete cds"		33203.53	39067.7
Hs.3712	AA448184	UBIQUINOL-CYTOCHROME C REDUCTASE IRON-SULFUR SUBUNIT PRECURSOR		33254.57	38355.42
Hs.49007	AA100296	H.sapiens PAP mRNA		34691.9	35581.06
Hs.28823	R37519	"Homo sapiens neuropilin-2(a17) mRNA, complete cds"		36256.89	33037.64
Hs.82159	R27585	Proteasome component C2	PSMA2	33242.18	35303.71
Hs.76845	W05628	Homo sapiens mRNA for L-3-phosphoserine-phosphatase homologue		35927.83	31553.12
Hs.7957	AA600189	Double-stranded RNA adenosine deaminase	ADAR	35979.14	30984.96
Hs.1369	R09561	"Decay accelerating factor for complement (CD55, Cromer blood group system)"	DAF	28280.45	38148.5
RG.51	W86100			28529.21	37640.83
Hs.80475	AA460830	"Homo sapiens (clone mf.18) RNA		32949.78	33104.42

UniGene	Acc	Gene Description	Gene	LG	HG
		polymerase II mRNA, complete cds"			
Hs.2064	AA487812	Vimentin	VIM	30411.88	35468.75
Hs.89137	AA464566	Human mRNA for LDL-receptor related protein		29810.95	35825.83
Hs.7763	AA504342	"Human mRNA for KIAA0256 gene, complete cds"		28667.1	36770.34
Hs.89801	AA481547	H.sapiens mRNA for LPAP protein		27142.43	38243.19
Hs.273	W85914	Galactocerebrosidase	GALC	29292.01	35239.53
Hs.808	AA490991	"Homo sapiens HnRNP F protein mRNA, complete cds"		26014.7	37572.27
Hs.75822	AA446222	"Human mRNA for KIAA0216 gene, complete cds"		32515.44	30981.18
Hs.24025	R26337	"ESTs, Weakly similar to C06A6.3 gene product [C.elegans]"		26143.37	36723.68
Hs.78941	AA436591	"Human cellular proto-oncogene (c-mer) mRNA, complete cds"		31117.7	30413.1
Hs.34606	AA402960	"Human HLA class III region containing NOTCH4 gene, partial sequence, homeobox PBX2 (HPBX) gene, receptor for advanced glycosylation end products (RAGE) gene, complete cds, and 6 unidentified cds"		31532.13	29282.97
Hs.88778	AA280924	Carbonyl reductase	CBR	28080.19	32476.47
Hs.84123	AA485539	"Human mRNA for KIAA0365 gene, partial cds"		29380.54	31127.86
Hs.73799	AA490256	Alternative guanine nucleotide-binding regulatory protein (G) alpha-inhibitory-subunit	GNAI1	28072.43	31709.99
Hs.1540	AA129338	"Human (clone N5-4) protein p84 mRNA, complete cds"		28556.03	30634.62
Hs.94466	AA448487	H.sapiens hPTPA mRNA		29534.53	28990.42
Hs.76507	AA625666	"Homo sapiens Pig7 (PIG7) mRNA, complete cds"		32284.21	26212.44
Hs.75889	AA460599	V-jun avian sarcoma virus 17 oncogene homolog	JUN	27101.04	30855.16
Hs.42957	AA422058	H.sapiens mRNA for D1075-like gene		30279.34	27499.56
Hs.38022	N91307	ESTs		26401.21	30394.82
Hs.8653	AA490300	"Human PDGF associated protein mRNA, complete cds"		27568	28912.59
Hs.76293	AA486085	THYMOSIN BETA-10		25142.41	30457.52
Hs.47831	N76599	ESTs		24890.33	30641.2
Hs.83758	AA397813	CDC28 protein kinase 2	CKS2	27805.31	27613.93
Hs.107171	R71531	ESTs		28065.79	26573.28
Hs.89751	N91385	CD20 RECEPTOR		26008.2	27177.65
Hs.107942	W76032	"ESTs, Highly similar to HYPOTHETICAL 30.5 KD PROTEIN C30A5.3 IN CHROMOSOME III [Caenorhabditis elegans]"		26399.4	26586.85
Hs.75305	AA455316	"Homo sapiens immunophilin homolog ARA9 mRNA, complete cds"		27594.14	24511.31
Hs.1447	AA069414	Glial fibrillary acidic protein	GFAP	24365.91	27617.42
Hs.3548	AA029842	"H.sapiens MTCP1 gene, exons 2A to 7 (and joined mRNA)"		29248.4	22182.65
Hs.2890	R89715	"Protein kinase C, gamma"	PRKCG	23993.39	27074.05
Hs.5398	N59764	"Human guanosine 5'-monophosphate synthase mRNA, complete cds"		22973.33	27806.4
Hs.11465	AA441895	"Human glutathione-S-transferase homolog mRNA, complete cds"		23514.69	25986.39
Hs.118475	R91078	Cytochrome P450 IIIA7 (P450-HFLa)	CYP3A7	24577.65	24228.33
RG.35	N90246			22511.49	25486.64
Hs.13137	AA490771	H.sapiens mRNA; UV Radiation Resistance Associated Gene		23102.01	24308.85
Hs.47701	N73115	ESTs		21667.54	25372.87

UniGene	Acc	Gene Description	Gene	LG	HG
Hs.74085	AA397819	NKG2-D TYPE II INTEGRAL MEMBRANE PROTEIN		21095.7	25923.91
Hs.81728	AA457199	"Human retinal protein (HRG4) mRNA, complete cds"		19634.13	27071.1
Hs.52763	T67474	ESTs		23793.16	21409.2
Hs.75923	AA452374	Syntaxin 5A	STX5A	21330.49	23534.99
Hs.23119	AA455272	H.sapiens mRNA for ITBA1 protein		20477.13	23566.21
Hs.47433	N73201	ESTs		20147.7	23791.57
Hs.3436	R78607	"Homo sapiens doc-1 mRNA, complete cds"		20295.63	23375.23
Hs.89111	AA284568	"Human adult heart mRNA for neutral calponin, complete cds"		18468.56	24678.42
Hs.84640	W93317	ESTs		19526.01	22514.73
Hs.76194	AA456616	Ribosomal protein S5	RPS5	21213.93	20179.77
Hs.10716	T60223	"Ribonuclease L (2',5'-oligoisoadenylate synthetase-dependent)"	RNASEL	20068.41	20973.43
Hs.118220	W17246	"ESTs, Highly similar to PEPTIDYL-PROLYL CIS-TRANS ISOMERASE [Drosophila melanogaster]"		22990.64	17599.4
Hs.81352	H77855	ESTs		19900.95	20527.58
Hs.15744	W23931	"ESTs, Moderately similar to FceRI gamma-chain interacting protein SH2-B [R.norvegicus]"		19559.57	20009.91
Hs.82173	R79935	"Human TGF-beta inducible early protein (TIEG) mRNA, complete cds"		19549.88	19901.95
Hs.44899	N48103	ESTs		18463.06	20916.15
Hs.9629	AA488233	H.sapiens mRNA for prcc protein		17947.14	21376.25
Hs.67102	T98612	"Ferritin, light polypeptide"	COL1A1	19292.54	19348.49
Hs.89839	N90246	TYROSINE-PROTEIN KINASE RECEPTOR EPH PRECURSOR		18718.38	19790.27
Hs.2048	AA284528	"Protease, serine, 2 (trypsin 2)"	PRSS2	18363.49	19668.13
RG.41	W48713			17701.89	20313.28
Hs.78883	H23187	Carbonic anhydrase II	CA2	20720.62	16677.23
Hs.83611	AA232647	"Human mRNA for DB1, complete cds"		14721.85	22675.65
Hs.80324	AA521083	H.sapiens mRNA for protein phosphatase 6		17104.3	20185.62
Hs.85111	AA485626	S-adenosylhomocysteine hydrolase	AHCY	18579.31	18658.2
Hs.31439	AA459039	"Human Placental bikunin mRNA, complete cds"		18188.28	18683.29
Hs.84285	AA487197	Ubiquitin-conjugating enzyme E2I (homologous to yeast UBC9)	UBE2I	19352.75	17025.28
Hs.3005	AA284693	Transcription factor AP-4 (activating enhancer-binding protein 4)	TFAP4	17152.3	18483.96
Hs.96038	AA027840	H.sapiens mRNA for RIT protein		17718.88	17845.55
Hs.76155	AA464731	"Human mRNA for calgizzarin, complete cds"		15412.75	20012.37
Hs.76144	R56211	"Platelet-derived growth factor receptor, beta polypeptide"	PDGFRB	17695.33	17592.65
Hs.77060	AA070997	"Proteasome (prosome, macropain) subunit, beta type, 6"	PSMB6	18150.62	16812.42
Hs.79601	R69153	ESTs		21600.4	13212.39
Hs.74594	W73892	"Human putative tumor suppressor (LUCA15) mRNA, complete cds"		18031.18	16394.02
Hs.88859	AA281548	"Human putative holocytochrome c-type synthetase mRNA, complete cds"		16140.06	18110.99
Hs.64227	AA423944	"Human p37NB mRNA, complete cds"		19117.21	15077.09
Hs.74950	AA486082	H.sapiens mRNA for putative serine/threonine protein kinase		13574.92	20437.35
Hs.31389	R01638	"Homo sapiens mRNA for HYA22, complete cds"		16757.56	17019.03
Hs.5174	AA281137	Ribosomal protein S17	RPS17	15954.67	17077.21
Hs.7166	H05580	"PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR"		16531.69	16042.06

UniGene	Acc	Gene Description	Gene	LG	HG
Hs.82028	AA487034	"Transforming growth factor, beta receptor II (70-80kD)"	TGFBR2	13132.46	19066.27
Hs.3850	R94775	"ESTs, Weakly similar to HYPOTHETICAL 47.5 KD PROTEIN IN SECB-TDH INTERGENIC REGION [Escherichia coli]"		16690.49	14993.17
Hs.119571	T98612	Alpha-1 type 3 collagen	COL3A1	16205.37	15100.87
Hs.44782	T69926	"Myosin, heavy polypeptide 9, non-muscle"	MYH9	15652.12	15529.3
Hs.31053	AA504554	"Human cytoskeleton associated protein (CG22) mRNA, complete cds"		15066.05	15757.53
Hs.75355	AA490124	"Human epidermoid carcinoma mRNA for ubiquitin-conjugating enzyme E2 similar to Drosophila bendless gene product, complete cds"		14536.89	16186.46
Hs.108620	W52803	"ESTs, Highly similar to gene Fif protein [M.musculus]"		17722.88	12874.32
Hs.85200	R01669	"Human mRNA for protein disulfide isomerase-related protein P5, complete cds"		15882.02	14396.89
Hs.36218	W00987	ESTs		14180.02	15929.94
Hs.76476	AA487346	Cathepsin H	CTSH	14452.06	13744.68
Hs.75616	AA482324	"Human mRNA for KIAA0018 gene, complete cds"		13683.05	14255.38
Hs.72082	AA148736	"Syndecan 4 (amphiglycan, ryudocan)"	SDC4	13614.04	14272.64
Hs.7476	AA480826	"Human mRNA for proton-ATPase-like protein, complete cds"		11667.43	16129.12
Hs.75317	AA425612	"Human X-linked PEST-containing transporter (XPCT) mRNA, partial cds"		13418.03	14373.67
Hs.119619	R02609	"ESTs, Weakly similar to VACUOLAR ATP SYNTHASE SUBUNIT C [H.sapiens]"		17823.15	9922.45
Hs.106194	R32428	EST		11883.87	15601.89
Hs.83343	R43544	Ribosomal protein L32	RPL32	12431.78	14675.41
Hs.74136	R86304	"Homo sapiens clone 23689 mRNA, complete cds"		13336.36	13537.61
Hs.3378	AA504348	Topoisomerase (DNA) II alpha (170kD)	TOP2A	11368.51	15454.06
Hs.75132	AA485992	"Human IEF SSP 9502 mRNA, complete cds"		12572.65	14155.2
Hs.91370	AA490124	ESTs		12036.96	14631.51
Hs.822	N54596	"Human Krueppel-related zinc finger protein (H-plk) mRNA, complete cds"		13816.02	12109.81
Hs.94931	W88615	"Human mRNA for KIAA0314 gene, partial cds"		12173.97	13679.48
Hs.84898	AA598950	Cathepsin B	CTSB	11919.64	12746.35
Hs.953	AA452725	NUCLEOBINDIN PRECURSOR		13198.05	11152.58
Hs.79353	W33012	"Homo sapiens E2F-related transcription factor (DP-1) mRNA, complete cds"		9045.16	15177.23
Hs.75963	N54596	Insulin-like growth factor 2 (somatomedin A)	IGF2	12916.31	11304.11
Hs.75737	AA164439	"Human autoantigen pericentriol material 1 (PCM-1) mRNA, complete cds"		12398.9	11513.71
Hs.8724	AA521346	H.sapiens mRNA for Ndr protein kinase		11796.26	11993.78
Hs.23807	R25153	ESTs		13450.48	10067.89
Hs.89462	H05774	"Diacylglycerol kinase, gamma (90kD)"	DAGK3	12370.32	11078.95
Hs.2475	AA281057	"Human mRNA for KIAA0019 gene, complete cds"		11796.16	11330.56
Hs.75104	AA496837	"Human (clone E5.1) RNA-binding protein mRNA, complete cds"		10476.4	12562.82
Hs.75862	AA456439	"Human homozygous deletion target in pancreatic carcinoma (DPC4) mRNA, complete cds"		11623.19	11413.66
Hs.68731	AA404293	"Human triadin mRNA, complete cds"		11329.59	11297.98
Hs.79361	AA454743	"Human protease M mRNA, complete cds"		11018.01	11487.96

UniGene	Acc	Gene Description	Gene	LG	HG
Hs.75850	N59851	"Human mRNA for KIAA0269 gene, complete cds"		9668.76	12711.08
Hs.43509	AA029963	"Human ataxin-2 related protein mRNA, partial cds"		10077.98	12033.22
Hs.21729	T72698	H.sapiens mRNA for splicing factor SF3a120		9581.63	12410.7
Hs.5085	AA004759	"Homo sapiens dolichol monophosphate mannose synthase (DPM1) mRNA, partial cds"		11980.15	9608.32
Hs.709	H12903	Deoxycytidine kinase	DCK	9270.7	12169.99
Hs.36779	H53499	"ESTs, Moderately similar to hypothetical protein p18 [H.sapiens]"		9102.82	11187.11
Hs.743	H79353	"Human Fc-epsilon-receptor gamma-chain mRNA, complete cds"		10539.72	9074.7
Hs.17575	T95462	ESTs		11109.6	8455.28
Hs.64639	AA251800	"Human glioma pathogenesis-related protein (GliPR) mRNA, complete cds"		9227.21	10042.17
Hs.9930	R71440	Collagen-binding protein 1	CBP1	8146.85	10888.3
Hs.75175	H29077	Lysosome-associated membrane protein 1	LAMP1	9509.47	9491.64
Hs.75879	T68202	"Human DXS8237E mRNA, partial cds"		9285.53	9688.15
Hs.1940	AA504943	"Crystallin, alpha B"	CRYAB	8105.99	10477.75
Hs.7753	R78585	"ESTs, Highly similar to RETICULOCALBIN PRECURSOR [Mus musculus]"		9508.68	8960.29
Hs.75564	AA456183	"Homo sapiens mRNA for CD151, complete cds"		9588.54	8829.16