Université d'Ottawa • University of Ottawa

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	TRAULSEN, Kathryn E. A.
ADRESSE POSTALE / MAILING ADDRESS:	310-171 O'CONNOR STREET OTTAWA ON K2P1T4
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
M.Sc. (Microbiology spec.: Human and Molecular Genetics)	2003
TITRE DE LA THÈSE / TITLE OF THESIS: TOWARDS A TRANSCRIPT MAP OF 4Q34	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses successeurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

April 03/03.

DATE

K Traulsen

(AUTEUR) SIGNATURE (AUTHOR)

Université d'Ottawa • University of Ottawa

Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

TRAULSEN, Kathryn E.A.

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.Sc. (Microbiology and Immunoloy - spec.: Human and Molecular Genetics)

GRADE - DEGREE

Biochemistry, Microbiology and Immunology

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Towards A Transcript Map of 4q32-q34

Dennis Bulman

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

J. Bell

M. Holcik

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Towards A Transcript Map of 4q32-q34

By

Kathryn E. A. Traulsen

THESIS
Submitted to the School of Graduate Studies in partial fulfilment of the requirements for the degree of

Master of Science

Department of Biochemistry, Microbiology and Immunology,
Human Molecular Genetics
Faculty of Medicine
University of Ottawa

Abstract

With the completion of the sequencing of the human genome near, the next phase will be to identify and annotate all transcriptional units. We have utilized a procedure that directly selects cDNA from genomic DNA in order to isolate putative transcriptional units for identification of novel genes at the 4q34 locus. In this procedure, cDNA fragments were isolated following the hybridization of cDNA pools to 7 BAC clones spanning the 4q34 region. The 4q34 region is approximately 2.0 Mb and contains 7 known genes and 8 predicted genes. In addition, EST evidence annotated in the public databases supports the notion that there are additional transcriptional units in this region. The primary cDNA pool used in this procedure was generated with a universal primer for amplification and cloning. Approximately 350 clones were analyzed by automated fluorescence sequencing and 26 clones were shown to originate from DNA at the 4q34 locus. The other clones included rRNA, mitochondrial DNA, repetitive sequences and low-copy repeat elements, similar to the results obtained in comparable cDNA selection attempts. The tentative transcriptional units have been arranged on the current map of the 4q34 region. This map will provide insight into the organization and function of this chromosome, and provide the preliminary framework for a detailed transcription map of the region. Furthermore, novel gene identification at this locus will provide candidates for Parkinson's disease, which has been mapped to this region by our laboratory.

Acknowledgements

I would like to extend thanks to my supervisor, Dr. Dennis Bulman, who provided me with a great amount of support and guidance.

I would like to thank my advisory committee, Dr. Dave Picketts, and Dr. Ken Dimock for keeping me on track.

I owe my thanks to Dr. Johanna Rommens for her insight and all of the materials she provided to us.

Thanks to my parents for supporting me through all my years of university, I would not be here without them.

Finally, I would like to thank my fiancé, Rob, for his support and love, which has helped me reach my goals.

Table of Contents

List of Tables

List of Figures

List of Abbreviations[1]

A	Adenine
AC	Accession of a BAC (e.g. AC105285 which is RP11-10K16)
AK	Accession of a cDNA clone (e.g. AK024995)
ATP	Adenosine Triphosphate
BAC	Bacterial Artificial Chromosome
Bp	Base Pairs
BLAST	Basic Local Alignment Search Tool
BLASTp	Basic Local Alignment Search Tool, protein
C	Cytosine
°C	Degrees Celsius
cDNA	Complementary DNA
Ci	Curie
Contig	Contiguous DNA
cM	Centimorgan
dATP	Deoxyadenosine Triphosphate
dCTP	Deoxycytosine Triphosphate
dbEST	database of Expressed sequence tags
ddH_20	double-distilled water
dGTP	Deoxyguanine Triphosphate
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleoside Triphosphate
dTTP	Deoxythymidine Triphosphate
EDTA	Ethylenediamine Tetraacetic Acid
EST	Expressed Sequence Tag
FBX08	F-box only protein 8
FLJ	Annotation of proteins predicted by protein homology
FSHD	Facioscapulohumeral muscular dystrophy
G	Guanine
GALNT7	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 7
h	Hour(s)
HGP	Human Genome Project
HAND2	Heart-and-neuralcrest derivatives expressed 2
HLH	Helix-loop-helix domain
HMG2	High-mobility protein group 2

[1] Abbreviations referring to genes are italicized (e.g. *GALNT7*). When referring to a human gene, all letters are capitalized (e.g. *HMG2*). When referring to murine genes, only the first letter is capitalized.

HS	Hierarchical shotgun
HUGE	Human Unidentified Gene-Encoded Large Proteins database
IRD	Infrared
Kb	Kilobases
kDa	Kilodaltons
KIAA	Accession for genes from the HUGE database (e.g. KIAA1712)
L1	*Kpn1* fragment, LINE sequence
LB	Luria-Broth
LINE	Long interspersed repeat elements
LOC	Accession for an *in silico* predicted protein (e.g. LOC256573)
LOD	Logarithm of the odds
M, mM	Molar (moles per litre), Millimolar
Mb	Megabase
MCS	Multiple Cloning Site
min.	minute
mL	milliliter
mRNA	messenger RNA
MW	Molecular Weight
N, n	any nucleoside base (A,T,G or C)
NA	Not applicable
NCBI	National Centre for Biotechnology Information
nL	Nanolitre
NPYR1	Neuropeptide Y receptor1 (also 3 and 5)
nr	Non-redundant database
NT	genomic contig annotated in NCBI (e.g. NT_006257)
O/N	Overnight
OMIM	Online Mendelian Inheritance in Man
ORF	Open Reading Frame
Pax5	Paired Box Gene 5
PCR	Polymerase Chain Reaction
PD	Parkinson's Disease
RNA	Ribonucleic Acid
RNaseA	Ribonuclease A
RNaseH	Ribonuclease H
rRNA	Ribosomal Ribonucleic Acid

RPI1	Roswell Park (Cancer) Institute library[2]
RT-PCR	Reverse Transcriptase - Polymerase Chain Reaction
RXG	Amplification and cloning primers
RXGKS/SK	Intake primers for pksBluescript Library
s	Seconds
SAP30	Sin3 associated protein 30
SCFs	E3 ubiquitin protein ligases
SCRG1	Scrapie-responsive protein 1
SDS	Sodium Dodecyl Sulfate
SINE	Short interspersed repeat elements
SN	Substantia Nigra (pars compacta)
SSC	Saline Sodium Citrate
STS	Sequence tag site
T	Thymine
TE	Tris-EDTA
TIGR	The Institute for Genomic Research
Tris	Tris(hydroxymethyl) aminomethane
tRNA	Transfer Ribonucleic Acid
TSE	Transmissible spongiform encephalopathy
U	Units
UCH-L1	Ubiquitin ligase L1
µCi	Microcurie
µL	Microlitre
UTR	Untranslated region
UV	Ultraviolet
V	Volts
VEGF	Vascular Endothelial Growth Factor
WGS	Whole genome shotgun
www	World wide web
XM	Accession for mRNA (e.g. XM_003527 represents *GALNT7*)
YAC	Yeast Artificial Chromosome
YY1	Transcription factor YY1

[2] The original BAC library was labelled RPCI-1 for Roswell Park Cancer Institute-library, but was shortened to RPI1. In the text when referring to a BAC clone from this library, only the numeric portion is included (e.g. RPI1-694 K14 = 694 K14).

Chapter 1

1.0 Introduction

1.1.1 Towards a Draft Sequence of the Human Genome

Sequencing of the human genome was predicated on two major technical advances, a means to manipulate large DNA fragments and an efficient and cost effective method of sequencing DNA. The advent of this technology was the true catalyst for the expanding depth and scope of our knowledge of macromolecular sequences [1].

Sanger and Coulsen originally accomplished the dideoxy-sequencing of DNA in early 1977 [2,3]. Modifications to this method by Hood *et al.* [4] permitted DNA to be sequentially read by computers. Automated DNA sequencers became available shortly there after [5]. The use of a shotgun approach for sequencing genomes was proposed in 1981 by Anderson [6] and became widely utilized quickly [7,8,9].

Shotgun sequencing involves digestion of a large DNA fragment into small fragments, which are individually sequenced. Determining the overlaps between the smaller fragments and piecing the sequence together can generate the complete sequence of the large fragment. Currently, shotgun sequencing is highly automated and the assembly of simple genomes, which have a low occurrence of repetitive elements, such as viruses, organelles and bacteria has been quite successful with this technique. Cloning biases and failures in the sequencing chemistry dictate that generally the random data generated from shotgun sequencing is not sufficient to yield the complete sequence. Instead, scientists rely on supplementary data like previously characterized genetic and

physical maps to fill in the sequencing gaps. Additionally, sequencing of more complex genomes such as *Drosophila* (3 % repetitive) and human (greater than 50 % repetitive) has been aided with the use of a clone-based strategy [10]. In February 2001, the Human Genome Project (HGP) [11] and Celera Genomics [12] each published their version of the draft human genome sequence.

The HGP, a core group of 16 laboratories worldwide, based their sequencing strategy on a hierarchical shotgun (HS) approach. The first step taken in HS sequencing was to generate a physical map of the human genome [13]. A number of genomic libraries were constructed using Bacterial Artificial Chromosomes (BACs) as the vector [13,14,15]. The size of the genomic DNA fragments cloned into the BACs ranged from 75 – 350 kb. Overlapping clones encompassing large regions of contiguous DNA (contigs) were first organized by chromosome in a process called "fingerprinting" [16]. This involves digesting the individual BACs with a restriction enzyme, *Hind III* was initially used, and determining fragment sizes by agarose gel electrophoresis. The pattern of restriction fragments for each BAC provides a "fingerprint" allowing each BAC to be distinguished and the degree of overlap between BACs to be assessed, prior to sequencing each BAC. The restriction fragment fingerprints were used to assemble clone contigs and then known genetic markers, such as STSs, were evaluated as landmarks to ensure that overlapping clones were not misassembled. This was followed by the shotgun sequencing of the individual BAC clones, merging data from each clone and ultimately piecing together the entire genome [16]. A HS approach requires significant preliminary input of money and labour, but the sequence assembly is anchored to the genome by the physical map reducing the majority of the problems presented by repetitive and

duplicated elements [11]. Problems such as rearrangements may occur in some large-insert clones, however these are rectified and limited by sufficient mapping density (fold-coverage) and clone fingerprinting [11]. Another problem encountered was that certain DNA sequences, likely highly repetitive elements, cannot be propagated in bacteria or yeast [16].

In contrast, Celera Genomics, a biotechnology firm, chose to use a whole-genome shotgun (WGS) approach, which was initially proposed in 1997 [17], despite poor support [18,19]. WGS sequencing entails shotgun sequencing of the entire genome without assembling a physical map, and reassembling the entire collection to map the genome. Although the WGS approach avoids preliminary data collection, the risk of long-range misassembly is much higher. Because each sequenced component must be individually anchored to the genome, there are inherent problems when dealing with repetitive and duplicated elements [16]. Therefore, the assembly process may be elevated in cost, thus obviating the savings from avoiding the physical mapping step. It is also important to note that Celera had access to the publicly available HGP physical mapping data when assembling their data. Therefore, their approach was actually a combination of both methods because their final product was in fact a merger of the two data sets, and not strictly a whole-genome approach [10].

Estimates at the time of the draft sequence publications suggested that the human genome would be sequenced in 'polished' form by spring 2003, only 50 years after the discovery of the structure of DNA [20]. Sequencing centres have recently achieved sequencing rates between 1000 and 2000 base pairs (bp) per second, 24 hours a day, 7 days a week with fewer than 1 error in 100000 bases [21,22]. Despite this, genome

centres are now faced with the major challenge of overcoming the problems of sequencing and assembling certain elements of the genome, such as repetitive sequences, telomeres and centromeres, and large segmental duplications. Some elements are difficult to clone into conventional vectors such as BACs, difficult to sequence and difficult to assemble into contigs because of their repetitive nature. Therefore, these regions of the genome are generally not represented in the current published shotgun sequence data sets [23].

1.1.2 The BAC physical map

The human genome sequence available in the public database was constructed with a HS sequencing approach. The primary DNA sequence available from the human genome project is based on the overlapping sequence of large-insert BAC clones constructed during the physical mapping portion of the HGP. Each BAC is designated as either draft or finished sequence, and this is updated as more sequencing data is added to the databases. Draft sequence is an assembly of contigs, which will contain several gaps. Gaps are regions on the physical map where contigs cannot be shown to overlap. These gaps could be within a BAC or between two BACs. The exact order and orientation of the contigs in a draft BAC is unknown [54]. A BAC designated as "finished" is contiguously sequenced with a high quality standard. The nucleotide call error rate is 0.01 % and there are no gaps in finished sequence [54]. The quality of the sequence of the BACs coving the physical map dictates the quality of the genomic sequence in that region.

In order for the sequence from a BAC to be considered "complete", the DNA sequence from a minimum tiling path encompassing 8-10 fold coverage is required [54]. When one BAC has been sequenced completely, the tiling path of other BACs covering that region is minimized. In other words, any BACs that do not contribute unique sequence to the genomic region, for example a smaller BAC with sequence represented partially or completely by another BAC, are removed or reduced in size in the database in order to reduce the redundancy of, and to compress the information contained within the database. Therefore, in the end there remains only a single tiling path with overlapping fragments connecting individual BACs to produce the least amount of redundancy possible, while still providing complete and accurate coverage of the genomic region. Problems such as repetitive elements, duplications and sequences that cannot be propagated in BACs contribute to the difficultly in reconstructing the complete human genome sequence.

1.1.3 The repetitive problem

In the human genome, coding sequences comprise approximately 1.1 – 1.5 % of the genome, while repeat sequences account for at least 50 %, and likely much more [11,12,24]. The discovery of several classes of repetitive elements within the genome revolutionized the understanding of eucharistic genome organization [25]. There are five classes of repeats in the human genome: (1) transposon-derived repeats, referred to as interspersed repeats; (2) inactive retroposed copies of cellular genes usually referred to as processed pseudogenes; (3) simple sequence repeats, direct repetitions of relatively short fragments of DNA such as $(A)_n$, $(CA)_n$ or $(CGG)_n$; (4) segmental duplications, blocks of

10- 300 kilo bases (kb) copied from one region of the genome to another; and (5) blocks of tandemly repeated sequences which occur in telomeres, centromeres and ribosomal gene clusters [11]. Evidence suggests that the various types of repeats play important roles in chromosomal structure and replication [24].

Short interspersed repeated sequences (SINES) and long interspersed repeated sequences (LINES) are highly represented in the human genome. The major human SINE is the Alu sequence family, and the major LINE is the LI (KpnI) sequence. It was thought that understanding the function of these elements would follow the mapping of their genomic distribution [29]. However, the Alu family of repeats has been shown to dominate the poorly staining, gene rich regions of the genome, while the L1 family dominates the strongly staining gene-poor regions. These sequences may comprise up to 18% of the total DNA in a chromosomal band, and it has been proposed that this alone might account for the chromosomal banding patterns [25].

The focus of the sequencing effort is on the gene-rich, euchromatic DNA and estimates of the completeness of the sequencing effort exclude the seemingly "useless" heterochromatic regions of the chromosomes [11]. As a consequence, despite the fact that several chromosomes including 20, 21, 22, and Y have been sequenced to "finished" status [22], the sequence is not 100 % completed. The final 5 % of each chromosome has proven to be quite difficult to bring to "completed" status. Typical "finished" chromosomes still contain in the range of 10 contigs with gaps of various lengths between them, and are considered to be "operationally complete sequence of the euchromatic portion" of the genome [26].

Furthermore, assembly of contigs in the presence of repeats and duplications has proven to be one of the major rate-limiting factors in the completion of the human genome sequence. Assumptions that all repeats in the human genome are randomly distributed and members of known families, and thus detectable is not supported [18]. Alu sequences appear to occur in clusters and there are numerous regions of the genome that have undergone duplications. These duplicated regions are not classified members of repeat families. Clusters of repetitive elements and duplicated regions can result in the compression of the sequence when assembly is undertaken [11]. The initial sequencing efforts may have collapsed copies of repetitive elements into one sub-assembly. However, these sub-assemblies are easily identified by observing their level of coverage, which will be much higher than the average coverage depth [12].

New techniques will ultimately need to be devised to close these gaps and sequence the heterochromatic portions of the human genome [11,16,23,26]. It is likely that as our knowledge of the regulatory elements of the genome and the nature of these "junk" sequences improves, the value of understanding these gene poor regions of the genome will increase [27]. Regions like the centromeres and telomeres, although rich in repeats and duplicated segments, can be hot-spots for rapidly evolving genes, and are implicated in some two dozen diseases including Prader-Willi syndrome and DiGeorge syndrome [23]. These regions also provide crucial clues about evolutionary events and forces and the processes of mutation and selection [11]. Ideally, all heterochromatic regions will be sequenced [11]. However, this will require specialized data collection strategies and significant data editing, for which new strategies must be devised.

1.1.4 Genomic landscape and composition - Gene density

The genomic landscape of human DNA is non-uniform in the distribution of both repetitive elements and genes. Metaphase banding initially alluded to the variation in genomic composition when it was first described in 1970 [28]. Research into the composition of the various regions of the chromosomes, dark-staining and light-staining with Giemsa for example, has uncovered a relationship between functional and biochemical attributes. Euchromatic DNA, which stains poorly with Giemsa, is guanosine and cytosine (G/C) rich as well as gene-rich. In contrast, heterochromatic DNA stains strongly with Giemsa, and is adenine and thymine (A/T) rich and gene-poor. These relationships were known as early as 1978 [29].

Studies have also been undertaken to describe chromosomes at the molecular level. The density of genes is generally greater in regions with high G/C content, although many genes are located in G/C poor regions [11,12]. Euchromatin has been classified at the molecular level as long stretches of DNA differing in base composition, termed isochores. The isochores are denoted L, H1, H2 and H3 [30,31]. L isochores are defined as G/C poor, containing <43 % G/C, whereas the H isochors fall into 3 G/C-rich categories, containing at least 43 % (H1/H2) and greater than 48 % (H3) G/C [11,12]. Isochores H2 and H3 represent 8 and 5 % of the genome respectively and are 20-fold more enriched in genes when compared to L isochores [32].

Therefore, it can be seen that both genome landscape and molecular composition illustrate the non-uniform distribution of genes in the human genome. Venter *et al.* [12] define a gene desert (gene poor region) as being >500 kb of contiguous sequence without a gene. Under this definition, they find 605 Mega bases (Mb), approximately 20 % of the

human genome, in gene deserts. Interestingly, they also found that approximately 80 % of the genome is represented in the G/C-rich euchromatic cytogenetic bands.

1.1.5 Genomic landscape and composition - Gene number

As the human genome comes closer to being completely sequenced, the uncertainty of the total number of human genes seems to rise [33]. Estimates range from 28000 - 120000 [44,47,48,49,34]. The two recent publications of the human genome by the HGP and Celera narrow that range to between 26000 – 40000 [11,12]. Evidence based on extrapolations from EST, CpG islands and transcript density gives varying results [12]. Current estimates, such as that extrapolated from the chromosome 22 sequencing data, do not take all computationally predicted genes into account, indicating that the true number of genes might be higher [26]. Therefore, estimates of the total number of genes in the genome will increase significantly if it can be demonstrated that the computationally predicted genes represent transcriptional units [47]. Additionally, only in the range of 80 - 92 % of known genes can be mapped onto the current draft of the human genome, indicating that unsequenced regions will contribute substantially to gene numbers [11,36].

1.2 Decoding the Human Genome

The value of the human genome sequence is clear, but from the sequence alone we are not able to deduce where each individual gene resides nor are we able to determine the function of that particular gene [12,22]. With the availability of the human

genome, increasing attention is being focused on identifying the complete set of genes for all mammals. Complete annotation of the genome requires the precise localization of all the genes and genomic elements and assigning them to their products and/or functions [35]. The dilution of the coding sequence of genes resulting from extensive splicing, the low density of exons, elusive gene expression and the high proportion of interspersed repetitive elements, makes identifying complete transcriptional units difficult [11,26,36]. Several methods have been used successfully to identify 22000 of the predicted ~35000 human genes that are now annotated in GenBank, a publicly available database of human genome sequence [37]. For the purposes of a brief over-view, these methods will be divided into two broad categories: **(1) Identification of genes by high-throughput screening**; and **(2) Identification of genes by positional cloning**. The strengths and weaknesses of each method need to be considered when choosing to establish a transcript map of a particular region of the genome. A brief review of these techniques follows.

1.3 Identification of genes by high-throughput screening

Two broad techniques have been used extensively for the high-throughput identification of novel transcriptional units: (1) The identification of expressed sequence tags (ESTs) has proven to be invaluable in the high-throughput identification of transcriptional units; and (2) Theoretical approaches that involve the use of computational gene-finding programs which have been developed to identify genes on a non-biased, high throughput scale.

1.3.1 Gene identification with expressed sequence tags

Long before the completion of the sequencing project, most human genes will be sequence-tagged and placed on various physical maps [38]. A "transcript map" based on these sequence-tagged genes provides an estimate of the total gene number. Sequence tagged sites (STSs) are unique, anonymous genomic DNA fragments which can be detected using the polymerase chain reaction (PCR). The initial purpose of STSs was to act as the scaffold upon which the BAC contigs were constructed. In 1991, Sikela and coworkers proposed developing STSs from the 3` untranslated region (UTR) of mRNAs [39]. Later, these became known as expressed sequence tags (ESTs) and were developed by a number of groups, primarily as a means for gene identification [40,41,42]. The impetus for developing ESTs came from a number of laboratories, which began to sequence all of the clones in various complementary DNA (cDNA) libraries. ESTs provide information similar to STSs in that they can act as a scaffold for physical mapping, but they have the added benefit of providing the location of a transcribed sequence.

There are some limitations concerning ESTs, for example, it is widely recognized that EST databases may contain as much as 5 – 10 % artefact sequences such as intronic or intergenic DNA [43,44]. Furthermore, the EST databases have been estimated to contain only about 80 % of all human genes. The remaining 20 % of these genes likely represent tissue specific and low-abundance transcripts [45,46,47]. Das *et al.* have shown that tissue-restricted transcripts may have very few matching ESTs, indicating that such genes are probably not being effectively detected by EST sequencing [47]. However,

relief from some of these problems improves with the advancement of the human genome sequence and its respective database mining tools.

To date, 3000000 EST sequences are available [46] and annotated in databases such as Genemap '99 [47], Unigene and dbEST [48]. Because ESTs are theoretically generated from expressed sequences, transcriptional units are identified through the identification of ESTs. In general, each EST represents only a portion of a given transcriptional unit, illustrated by the fact that there are approximately 10 times more annotated ESTs than genes in the human genome. Therefore, analysis of clusters of ESTs using EST sequences to mine databases of cDNA libraries is necessary to identify full transcriptional units. The Institute for Genomic Research (TIGR) Gene Indices [49] attempts to identify transcriptional units represented by the EST data using computer algorithms.

Gene identification through ESTs has proven to be invaluable for the unbiased, high-throughput identification of genes within the human genome. Furthermore it has provided direction to the sequencing effort, an estimate of the total number of genes in the human genome and tools for researchers looking for novel genes. In the case of high-resolution mapping on a smaller scale, for example a particular chromosomal region or a candidate region for a disease causing mutation, the plethora of information from these databases is invaluable, but needs to be supplemented with experimental data.

1.3.2 Computational Gene Finding

Due to the fact that experimental methods for gene discovery are often time consuming and costly, and with the availability of the genome sequence, various

computational methods have been developed that can predict the locations of genes [50]. Since the 1990's, computer programs have been developed that integrate heterogeneous information about genes to locate them on the genome and predict their function. Factors taken into consideration include DNA signals involved in gene specification, statistical regularities that are characteristic of protein-coding sequences and similarity to known coding sequences [51]. Two basic approaches have been established for computational gene prediction: *sequence alignment method* and the integrated compositional and signal search or *ab initio gene prediction* [50].

The *sequence alignment method* is well established with considerable success. This type of program relies on the detection of similarity between a characterized sequence from a public database and an uncharacterized sequence of interest. If the sequence of interest demonstrates significant similarity to a sequence in the database, it can be suggested that they are from common evolutionary origin. By comparing unknown sequence to annotated DNA, protein, EST sequences or known sequence motifs, the structure or function of the uncharacterized sequence can be inferred, and subsequently evaluated experimentally.

The *sequence alignment* search has been proven to be useful in many cases where genes are orthologues (genes which are evolutionarily conserved amongst species) or paralogues (members of gene families within a species). However, the technique falls short when it comes to discovering genes that have no identifiable homologs in current databases. It has been suggested that only half of the vertebrate genes may be discovered using the sequence alignment method across phyla [26,52]. A further shortcoming of this method is due to the fact that sequence motifs and EST similarities will not elucidate the

entire structure of the predicted gene. These give only a similarity for a portion of the gene and confirmation of the entire gene structure will likely need to be explored experimentally. The Basic Local Alignment Search Tool (BLAST) family of search programs, available through the National Centre for Biotechnology Information (NCBI), provides a method to identify similarity between unknown sequences and the human genome, including known genes, proteins and conserved domains [53,54].

The *ab initio* method integrates coding statistics with signal detection into one framework. Coding statistics are measures of protein coding function, in other words statistical biases in DNA composition that are characteristic of coding regions. Coding function statistics are used in common gene-finding programs such as Genemark.hmm [55], Genomescan [56], and GRAIL [50] although there are many others. Signal sensors typically detect short sequences that are recognized by the machinery of the cell such as promoter elements, start and stop codons, splice sites, and poly-A sites. These sequences can be detected with various pattern recognition methods including simple consensus sequences, weight matrices or arrays, neural network and decision trees [50,51]. The template method takes both coding statistics and signal detection into consideration in order to generate significant predictive power [50]. These types of programs generally also integrate similarity with annotated sequence.

Developing computational software to predict genes is hindered by the organization of genes into exons representing coding sequence, and introns which are made up of what seems to be largely meaningless sequences that break up the coding sequences [12]. Current computational methods function poorly when it comes to the

identification of splice variants, non-protein-coding genes, and important regulatory regions like 5` and 3` untranslated regions, and core promoters [35].

While G/C content appears to have little effect on the accuracy of the current programs, exon length is significant to accuracy. Medium sized exons (70 – 200 bp) are the most consistently predicted while smaller and larger exons are more often missed or incorrectly predicted. Furthermore, first and last exons appear to be the most difficult to predict, possibly due to the weakness of detection of start and stop codons. No programs have yet been designed which accurately address complex genome organization including nested and overlapping genes and alternative splicing. As more sequence becomes available and our understanding of the sequence evolves, new computer algorithms will need to be devised to accommodate the increase in sequence information [6].

Gene prediction programs are indispensable tools for the initial analysis of genomic DNA, however experimental confirmation is essential to validate the presence of a gene [26,36,50]. Biases in the optimization procedures for computational programs lead to biases in the types of genes that are predicted. Evidence shows that as many as 30 % of computationally predicted exons do not overlap with any experimentally identified exons, indicating that these "over predictions" may not truly be exons [26]. These values vary depending on the region of the DNA sequence used for analysis [26]. The same analysis indicates that while approximately 95 % of genes are at least partially predicted by *ab initio* methods, few transcriptional units are complete and greater than 20 % of exons are not predicted at all [26,36]. Finally, a challenge with gene prediction programs is that the human genome draft sequence is constantly being updated making it essential for computational programs to re-evaluate the same sequence on a regular basis in order

to get complete, accurate predictions [6]. As the sequence information for which the gene prediction programs are based continues to grow, so does the accuracy of predicting genes [50]. Current program performances are insufficient for a reliable automated annotation approach alone [36,33]. *Ab initio* gene prediction is most useful when modified and re-run often to accommodate new information, and when used in combination with other methods [12,26].

1.4 Identification of genes by Positional cloning

Positional cloning involves identifying genes with mutations, solely on their position in the genome, and was first shown to be successful in 1986 by Orkin and colleagues with the cloning of the X-linked gene for chronic granulomatous disease [57]. By 1995, the technique had been well established, as it had been used to identify approximately 42 genes [58]. With the ever-increasing availability of human transcriptome maps, true positional cloning is becoming obsolete. Positional cloning by definition refers to identifying a gene based entirely on map location. Presently however, the large amount of information in the human genome databases includes not only the map location of numerous genes, but function and expression profiles. This information can be scrutinized in order to identify candidates for a disease based not only on position, but also on phenotypes of similar known diseases or specific expression patterns [58]. For example, Marfan syndrome was mapped to chromosome 15q by linkage methods. An attractive biochemical candidate, fibrillin, was found in this region and quickly identified to have disease-causing mutations in affected individuals (reviewed in 58). In the order of hundreds of genes have been isolated with positional cloning [1].

Linkage mapping is widely used in the study of heritable traits and in positional cloning strategies [12]. Genetic mapping with polymorphic markers of a specific phenotype allows for the localization of a disease-causing gene to a candidate region, often in the range of 0.5–5 megabases [58]. Genes can be identified from a candidate region of the human genome by a variety of methods. Methods including cross-species conservation, CpG island identification, exon amplification, cDNA selection, and exon tracking have all been used to successfully identify important genes with respect to particular heritable diseases. It might also be necessary to search the region for novel transcripts, which is time consuming and difficult [59]. Comparatively speaking, the relative effectiveness of each particular method must take into consideration the context of the specific genomic region being analysed [60].

1.4.1 Cross-species conservation

It is assumed that if a particular DNA sequence is conserved throughout evolution, this sequence provides an important function. Therefore, if a sequence is shown to be conserved across several species (orthologous), there is an implication that it may represent part of a gene or a gene regulatory element. The importance of this assumption can be illustrated by the fact that through the human-mouse homology maps many orthologous genes have been identified. There are 921 human diseases that have been identified to have a homologous mouse gene [61]. Identification of syntenic regions between two species demonstrates that the regions are in fact evolutionarily conserved [11]. Recently, approximately 90 % of human and mouse genomes have been shown to lie in syntenic regions, and approximately 80 % of mouse genes have a single human

othologue [62]. Conserved segments have also been illustrated between humans and fish, fly and worms (reviewed in 11). The zooblot, as described by Monaco *et al.* [63], is one experimental technique used to identify orthologous DNA sequences. A Southern blot containing genomic DNA from several different species is hybridized to a cDNA fragment thought to represent an expressed sequence. If the probe anneals to genomic DNA from several species, it is evidence that the probe might represent part of an orthologous gene sequence and can then be further characterized.

1.4.2 CpG Islands

The dinucleotide CpG occurs only at one-fifth the frequency expected if CpG occurrence was completely random [11]. The majority of CpG islands that have been identified in the human genome are found at the 5` ends of genes [64], and approximately 60% of human genes are associated with CpG islands [65]. Investigation of this anomaly uncovered the importance of CpG islands. CpG islands consist of short, dispersed regions of unmethylated DNA with a high frequency of the dinucleotide CpG relative to that seen in the bulk of the genome [66]. Methyl-CpG is mutated by deamination, therefore non-methylated residues are less likely to lead to mutation. Approximately 45000 CpG islands have been identified in the human genome, somewhat less if GC-rich repeats are masked [11]. Therefore, CpG islands are useful in gene identification and isolation techniques, and are often used in combination with other techniques such as computational gene prediction programs. Bulk purification of CpG islands from whole genomes or from individual genes can be accomplished based on the unusual base composition and methylation status of CpG islands [67]. However, the availability of the draft genome

sequence has greatly reduced the need to identify CpG islands experimentally. The isolation of novel genes by this method has been shown to be successful by Rommens *et al.* [68]. A DNA segment containing several transcription units that was initially identified on the basis of its ability to detect conserved sequences in several animal species by DNA hybridization was subsequently characterized and shown to correspond to a portion of the gene for cystic fibrosis [68,69,70].

1.4.3 Vectors that identify exons through splicing events

Exon trapping [71] and exon amplification [72] are methods that can be used to identify exons from random pieces of cloned genomic DNA. These methods are based on the fundamental principle that exons flanked by functional 5` and 3` splice sites that are cloned into a plasmid at a site containing an intron, also flanked by splice sites, will be processed *in vivo*. The cloning strategy for both procedures uses an exon-trapping cassette including a retroviral intron that essentially identifies functional splice acceptor and splice donor sites in the DNA cloned into the vector cloning site. During a retroviral life cycle, non-viral genomic sequences are correctly spliced and can be recovered as cDNA copies of the introduced segment (for a more detailed description, see ref. 72). Therefore, if the cloned DNA sequence contains splice donor/acceptor sites, it will be correctly spliced and available as a cDNA clone in the mammalian cell host. This technique has proven to be successful as Church *et al.* isolated *Pax-5* and novel gene fragments from chromosome 9 using this technique [73].

There are some limitations to the exon trapping/amplification protocols. For example, single-exon genes will not contain splice acceptor or donor sites, and will

therefore not be recovered with this method. Some cryptic splicing may also produce "exons" that are not part of any true gene. Furthermore, if splicing events are regulated by cellular functions such as cell division or DNA replication, or limited in a tissue specific manner, correct splicing may not occur in the packaging cell lines. Finally, some sequences may not be efficiently propagated in the retroviral vectors; therefore the recovered cDNA sequences may not represent all possible exons from the original genomic insert [71]. Validation of potential exons can be accomplished by generating a PCR probe corresponding to the spliced product, and using it to screen Northern blots.

1.4.4 Gene identification using cDNAs

Complementary DNA (cDNA) libraries contain DNA clones that represent the expressed sequences from a particular tissue or cell type. These libraries are constructed by a series of enzyme-catalyzed reactions involving *in vitro* synthesis of a cDNA copy from mRNA, its subsequent conversion into a double-stranded cDNA duplex followed by insertion into a cloning vector [74]. Analysis of the cDNA clones by hybridization, PCR, or sequencing can result in the discovery of novel genes. Mapping back an entire cDNA library containing 10^6 clones would be too time consuming for the individual laboratory. For this reason, techniques have been developed to screen cDNA libraries or pools against DNA segments or large insert clones such as BACs, yeast artificial chromosomes (YACs) or cosmids in order to select for cDNA clones which map to a particular chromosomal location.

Similar techniques described by Parimoo *et al.* [75] and Rommens *et al.* [68] use YACs, BACs or cosmids from a specific chromosomal location in a hybridization

protocol to isolate specific cDNA clones. Large insert clones are chosen from a genomic region of interest; denatured and/or digested and immobilized on nylon membrane. Complementary DNA pools generated from tissues of interest are hybridized to the BACs and selected clones are eluted in water. Sub-cloning and characterization allows for identification of known and novel genes that are present in the region covered by the BACs or YACs used in the hybridization. This technique is valuable when looking for genes expressed in a particular tissue and within a specific genetic region because cDNA pools can be generated from tissues of interest and individual BACs or BAC contigs can be chosen that span the predetermined locus of a disease gene. This technique has been used successfully by Tambini *et al.* to identify the XRCC2 DNA repair gene [76].

This technique relies on hybridization of target DNA to the immobilized BAC DNA. This presents two problems; 1) hybridization is not 100 % specific, in other words, cDNAs with moderate similarity can also be selected; 2) both the BAC sequence and the cDNA pools will contain a certain level of repetitive sequences that will be preferentially isolated due to their higher concentration. The stringency of the hybridization and the washing steps can be adjusted to levels that will sufficiently reduce the amount of non-specific hybridization that is obtained. Preannealing of the cDNA with sonicated human placental DNA is designed to block repetitive DNA sequences in order to limit the number of repetitive sequences that are selected for during hybridization.

1.5 Gene identification summary

Many approaches can be used to identify genes that cause heritable diseases. With the advent of more accurate and quicker computational gene-finding programs, it is

hoped that a complete transcription map of the human genome can be established in the near future [59]. The availability of such a map would facilitate the initial identification of disease causing genes, and allow scientists to focus their efforts on understanding how and why the mutations cause the disease, and more importantly, how to develop effective means for treating, curing or preventing the disease [58].

1.6. Scope of Thesis

1.6.1 A dynamic region of chromosome 4: 4q32-q34

The chromosomal region 4q32-q34 is represented in GenBank by a dynamic map, one that has gaps and ambiguities, and is a relatively uncharacterized region of the human genome. However, the presence of vast numbers of ESTs and predicted genes in the 4q32-q34 interval indicates that there are still novel genes to be identified in this region. Discrepancies between the physical and genetic maps create difficulties in estimating the physical size of this region. At the initiation of this research, the physical size of 4q32-q34 could have been anywhere between 2 – 10 Mb. Therefore, 4q32-q34 could have conceivably contained up to 200 genes based on the average gene density which can be as high as 23 genes/Mb [12]. Several gaps between contigs, and ambiguously mapped markers compromised the quality of the map of 4q32-q34.

The gene density over the entire 4q32-q34 region varies from areas that have virtually no genes or ESTs to areas that are saturated with genes and ESTs. Genes in the region have been associated with several disorders including aspartylglucosaminuria [77], susceptibility to pancreatic cancer [78] and progressive external ophthalmoplegia with

mitochondrial DNA deletions [79]. Linkage analysis in our lab has shown that Autosomal Dominant Parkinson's disease (PD) also maps to this region.

In order to identify candidate genes for PD in our linkage region, an extensive physical map based on BACs was constructed in our laboratory. The technique chosen to identify novel transcriptional units in the 4q32-q34 region is direct selection, developed by Dr. J.M. Rommens [80]. It is hypothesised that there are novel transcriptional units in this region that might provide additional candidates for the disease-causing mutation for PD. Genes known to reside in the region represent positive controls, in that identification of transcriptional units from known genes ensures that the cDNA selection procedure was successful.

As already discussed, repetitive elements make up greater than 50 % of the human genome and particular repeats, such as Alu sequences, are often found in greater numbers in gene-rich regions of the genome. BACs are large-insert clones with an average size of 140 Kb. It is therefore likely that every BAC contains a significant amount of repetitive sequence. The nature of the sequence in the BAC clones chosen for cDNA selection has been shown to influence the percent of cDNA clones mapping to the originating BACs [80]. Furthermore, repetitive elements (such as ribosomal RNA, repeats in genes and UTRs, etc.) will also be present in the cDNA pools used for hybridization. Repetitive elements by nature will be highly selected for in a hybridization protocol due to their higher concentration in the DNA. In order to block the repetitive elements that will be present in the chosen BACs and cDNA, the cDNA is first preannealed with sonicated human placental DNA. Following subcloning of the selected cDNAs, screening either by hybridization with radio-labelled sonicated placental DNA or sequencing will allow for

the elimination of repetitive DNA clones that are isolated. Furthermore, analysing the sequence of each clone using a repeat masking program, such as Repeatmasker, will identify clones composed of repetitive sequences and they will not be characterized further.

In order to analyse the clones obtained in this study, we propose to utilize all available resources to increase the depth of the study. Computationally predicted genes annotated in the 4q32-q34 region will be scrutinized to determine their validity with experimental analyses such as Northern blotting and RT-PCR. Sequences of clones will be assessed using an array of public database mining tools (BLAST, Genscan, BLASTp, Repeatmasker, Unigene etc.) in combination with experimental procedures such as PCR, RT-PCR, Northern and Southern blotting. We will take full advantage of all available resources.

An anticipated problem of this technique is based on the fact that DNA hybridization is not necessarily 100 % specific. Sequences with moderate homology may hybridize and be selected with this protocol. However, clones that map to the region with moderate similarity might indicate the presence of orthologues or paralogues within the region. By analysing the score obtained for each clone using the sequence similarity search tools like BLAST, it will be possible to determine if the sequence matched exactly, or with moderate similarity to the sequence present in the 4q32-q34 region, indicating possible members of gene families.

A benefit of direct selection is that cDNAs from specific tissues can be used to identify candidate genes. In this screen for Parkinson's disease candidate genes, we have chosen a cDNA pool that includes adult and fetal brain cDNAs as well as placental and

testis cDNAs. The choice of tissues is based on the knowledge that the pathogenesis of PD involves mainly the substantia nigra pars compacta of the adult brain, and therefore the assumption that the disease-causing gene is expressed in brain is warranted. Although in some cases, limiting the cDNA pool might be deleterious, in this case we hope to minimize the screening procedure by only analysing clones that are expressed in relevant tissues. The addition of further cDNA libraries including those derived from placenta and testis provides a means to identify transcripts that are expressed in other tissues, because the construction of the transcript map is not to be limited to a brain transcript map. I address the caveat that in fact our cDNA selection protocol may miss transcripts that are tissue specific to organs or tissues not included in the selection process. However, thorough the *in silico* analysis of the current sequence data of this region, the identification of genes without considering tissue expression will be accomplished.

1.7 Summary and statement of purpose

The construction of an accurate transcript map was undertaken to address the unfinished nature of the 4q32-q34 region. **By combining physical and genetic mapping data, the published draft human genome and its respective databases, database mining tools, and experimental positional cloning methods, we hope to improve the quality of the map in the 4q32-q34 region.** Availability of a complete transcript map of this region will provide candidate genes for linked genetic disorders and aid in annotation of the draft sequence of the human genome.

Chapter 2

2.0 Materials and Methods

2.1.1 Oligonucleotides

Primers were designed using the primer3 software available on the Internet (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi), and were synthesized by Sigma Genosys (Oakville ON). The company provided primers as dried pellets of known mass. Primers were diluted in 10 mM Tris pH 8.3 and H_2O to a final working concentration of 17 mM. A complete list of all oligonucleotides utilized throughout the course of this project is outlined in Appendix 1.

2.1.2 Complementary DNA pool

The cDNA pools were generated based on the methodology developed by Dr. J. Rommens at the Hospital for Sick Children in Toronto, ON [80]. A schematic overview of this technique is outlined in Figure 2.1. Our cDNA pools were generated from cDNA libraries utilizing oligonucleotides designed within the vector sequence of the library in addition to the RXG oligonucleotide (Figure 2.1 and Appendix 1 for primer sequences). The library was a Custom Lambda ZAP® II Substantia nigra Library (Stratagene Cloning Systems, La Jolla CA). Following amplification of the cDNA from the library, the DNA was precipitated in the presence of 200 μg of sonicated human placental DNA with the addition of 0.1 volume of 3M NaOAC and 2.5 volumes of ethanol at room temperature

for 30 min. The procedure following precipitation is as described by Rommens *et al.* [80].

Additionally, a cDNA pool was generously provided by Dr. Johanna Rommens and prepared as outlined in Rommens *et al.* [80]. The cDNA pool contained 10 μg of cDNA derived from adult brain, fetal brain testis and placental DNA and 200 μg of sonicated human placental DNA in 0.3 M NaOAc and ethanol, and was stored at -20°C until it was used.

2.2 Preparation of BACs

2.2.1 Isolation of BAC DNA

BAC DNA clones covering the 4q32-q34 region were obtained from the Center for Applied Genomics at the Hospital for Sick Children (Toronto, ON) as LB stab cultures. Liquid media (LB-broth) and solid media (LB-agar) were prepared as described [81].

BAC clones were grown in 40 mLs of LB containing Chloramphenicol (12.5 μg/mL) followed by incubation for 16-18 hours at 37°C with shaking (225 rpm). 35 mL of culture was pelleted in a JA-25.50 rotor (Beckman) centrifuged at 10000 rpm for 5 min. The pellet was resuspended in 10 mL STE (100 mM NaCl, 10 mM Tris-Cl pH 8.0, and 1 mM EDTA pH 8.0) and pelleted as above. The pellet was then resuspended in 3 mL of TGE (25 mM Tris-Cl pH 8.0, 50 mM Glucose, and 10 mM EDTA pH 8.0) with a glass pipette to limit foaming. Following the addition of 6 mL of freshly prepared alkaline lysis solution (0.2 M NaOH and 1% SDS) the mixture was inverted 4 times and incubated on ice for 10 min. The addition of 4.5 mL of ice cold 7.5 M NH_4OAc pH 7.6

Figure 2.1. The creation of a cDNA pool from total RNA **A.** The sequence of the oligonucleotide used for cDNA generation is shown. The underlined sequence is the *EcoR1* digestion site for sub-cloning; the bold and underlined sequence refers to the 3 base pair tag, which identifies the RNA source from which the cDNA was originally amplified. Tags identifying RNA sources used in this screen are also outlined. **B.** A generalized outline of the procedure used to generate the cDNA pool. The cDNA synthesis was designed to generate random RNA mixtures and did not require the ligation of oligonucleotide linkers at any step. First strand synthesis was prepared with an oligonucleotide that directed random priming (N_6) with reverse transcriptase that identified the tissue with the 3 base pair tag and directed subsequence expansion of the cDNA mixture (RXG). Second strand synthesis with Taq polymerase was facilitated with the addition of dATP by terminal transferase. Double stranded cDNA was amplified with limited rounds of PCR to expand the primary cDNA source for hybridization.

A.

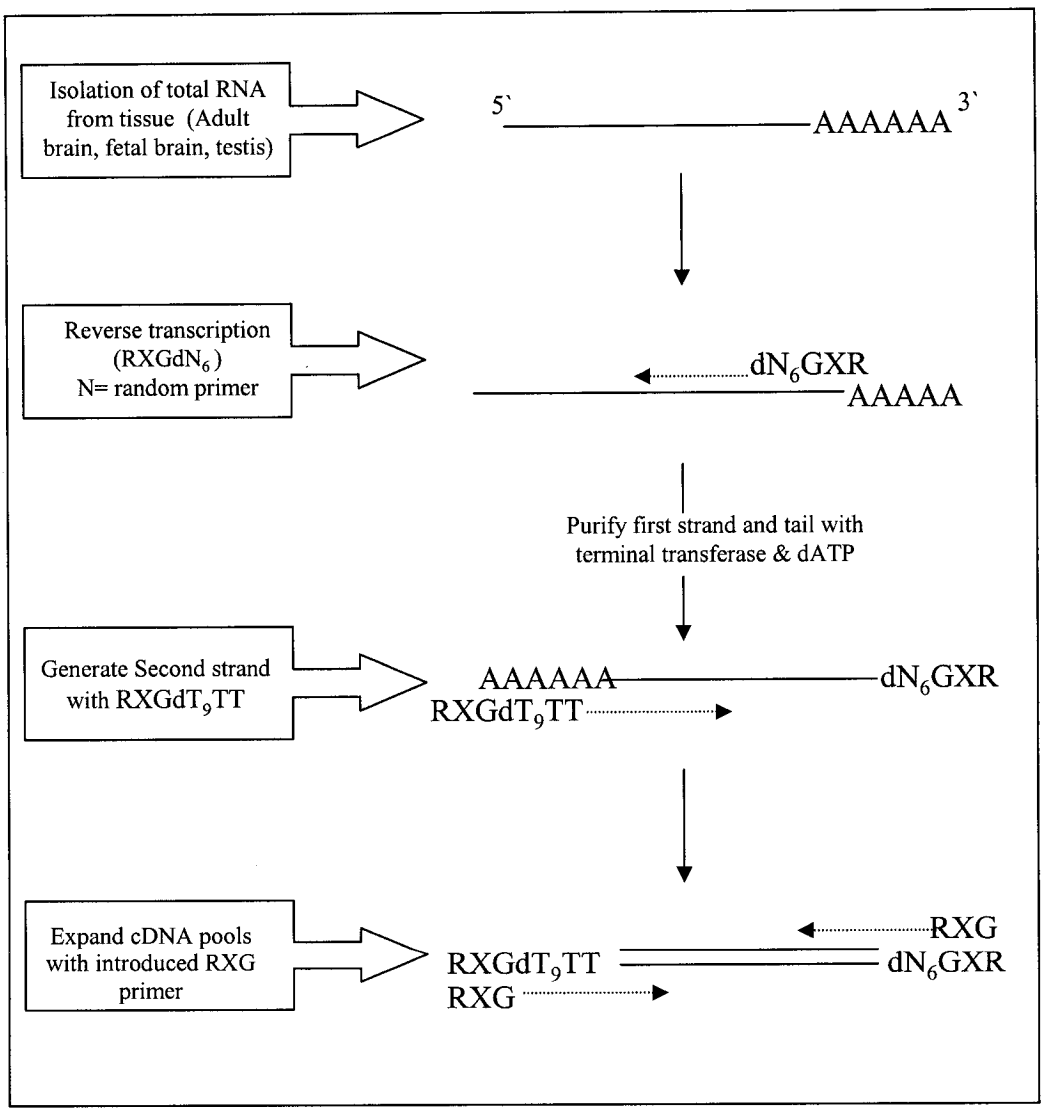
The RXG oligo sequence

CGGAATTCTCGAGATCT**A`B`C`**N_6 = $RXGdN_6$

A'B'C' = Tissue Tags

CAT – fetal brain **CGT** – adult brain **AGC** – testis **GCA** - placenta

B.



was followed by manual inversion 3 times and incubation on ice for 10 min. Centrifugation to pellet debris was carried out at room temperature in a JA-25.50 rotor (Beckman) for 20 min. at 15000 rpm. If the supernatant was not clear, additional centrifugation was required under the same conditions. The clear supernatant was transferred to a 50 mL Oak Ridge tube, and 0.6 volumes of isopropanol was added followed by 5 times manual inversion. The tubes were left to incubate at room temperature for at least 15 min. After centrifugation at 15000 rpm (Beckman) for 20 min., the supernatant was removed and the DNA pellet was air dried and resuspended in 525 µl of 2 M NH_4OAc pH 7.4 by repeated manual pipetting. The DNA suspension was transferred to a 1.5 mL microcentrifuge tube and vortexed briefly, then left on ice for 10 min. Following centrifugation at 14000 rpm (Eppendorf centrifuge 5417R) the supernatant was transferred to a clean microcentrifuge tube and an equal volume of isopropanol was added, the tube was manually inverted 5 times and left at room temperature for 10 min. The DNA was pelleted at 12000 rpm (Eppendorf) for 10 min. at room temperature, the supernatant was removed and the pellet was resuspended in 300 µl of 0.3 M NaOAc, pH 7.2. Following the addition of 750 µl of ethanol, the tube was manually inverted 5 times and left at room temperature for 30 min., or, alternatively, left over night at –20°C. The DNA was pelleted at 12000 rpm (Eppendorf) for 10 min., the supernatant was removed and the pellet was washed with 1 mL of 70 % ethanol. A final 10 min. centrifugation at 12000 rpm (Eppendorf) collected the pellet, which was air-dried. Resuspension of the pellet in 40 µl ddH_2O yielded between 200 – 500 µg of DNA.

2.2.2 Dot Blotting of BAC DNA

BACs were divided into pools of approximately five, depending on whether their inserts overlapped. Approximately 1 μg of each BAC was dot-blotted and immobilized on hybond-N nylon membrane (Amersham Pharmacia Biotech Inc., Baie d'Urfe, QB) following the manufacturers protocol. Briefly, the DNA was denatured with 0.4 M NaOH at 100°C for 10 min. and neutralized with cold 2 M ammonium acetate pH 7.0. Samples were applied to each well of the dot blot apparatus and a vacuum was used to draw the liquid through the membrane. Each well was washed with 2XSSC (0.3 M NaCl, 30 mM $Na_3C_6H_5O_7 \bullet 2H_2O$, pH 7.0), and the apparatus was disassembled. Finally, the entire membrane was washed by emersion in 2XSSC followed first by air drying then UV cross linking with 120 000 $\mu J/cm^2$ UV in a CL 1000 Ultraviolet Crosslinker (Ultra-Violet Products Inc., Cambridge England). Membranes were stored dry at room temperature until required.

2.2.3 Southern Blotting of BAC DNA

BAC DNA was linearized by restriction digest with 10 Units (U) of *Not I* (Amersham) at 37°C for 2 hours in a standard reaction. The linearized BACs (between 5–10) were electrophoresed for 1 – 2 hours on a 1% agarose gel in preparation for Southern blotting. The gel was immersed in 1.0 M HCl for approximately 10 min. The gel was then transferred to denaturizer (0.5M NaOH, 1.5M NaCl) for 2 times 15 min. DNA fragments were then transferred by capillary action to Hybond-N nylon membrane (Amersham) as previously described [82]. The membrane was neutralized by soaking

twice for 15 min in neutralizer (1 M Tris-Cl pH 7.5 and 1.5 M NaCl), air-dried and UV cross-linked as outlined above. Excess membrane was trimmed to minimize hybridization volumes. Duplicate membranes were made concurrently in order to have one blot for direct selection and one blot for the subsequent characterization of selected clones. Membranes were stored dry at room temperature until use.

2.3 Hybridization of cDNAs to immobilized BAC DNA

2.3.1 Preparation of cDNA for hybridization

The prepared cDNA pool (library amplified or from Dr. J. Rommens) and 200 μg of sonicated human placental DNA were precipitated at 14000 rpm (Eppendorf) for 15 min. at 4°C. The pellet was resuspended in 100 μl 0.2 M NaOH, 1mM EDTA and heated to 60° C to ensure complete denaturation. The addition of 1/10 volume of 2 M NH4OAc, pH 5.4 neutralized the solution and 1/4 volume 20XSSC was added and the cDNA was preannealed with human placental DNA at 55°C for 1 hour.

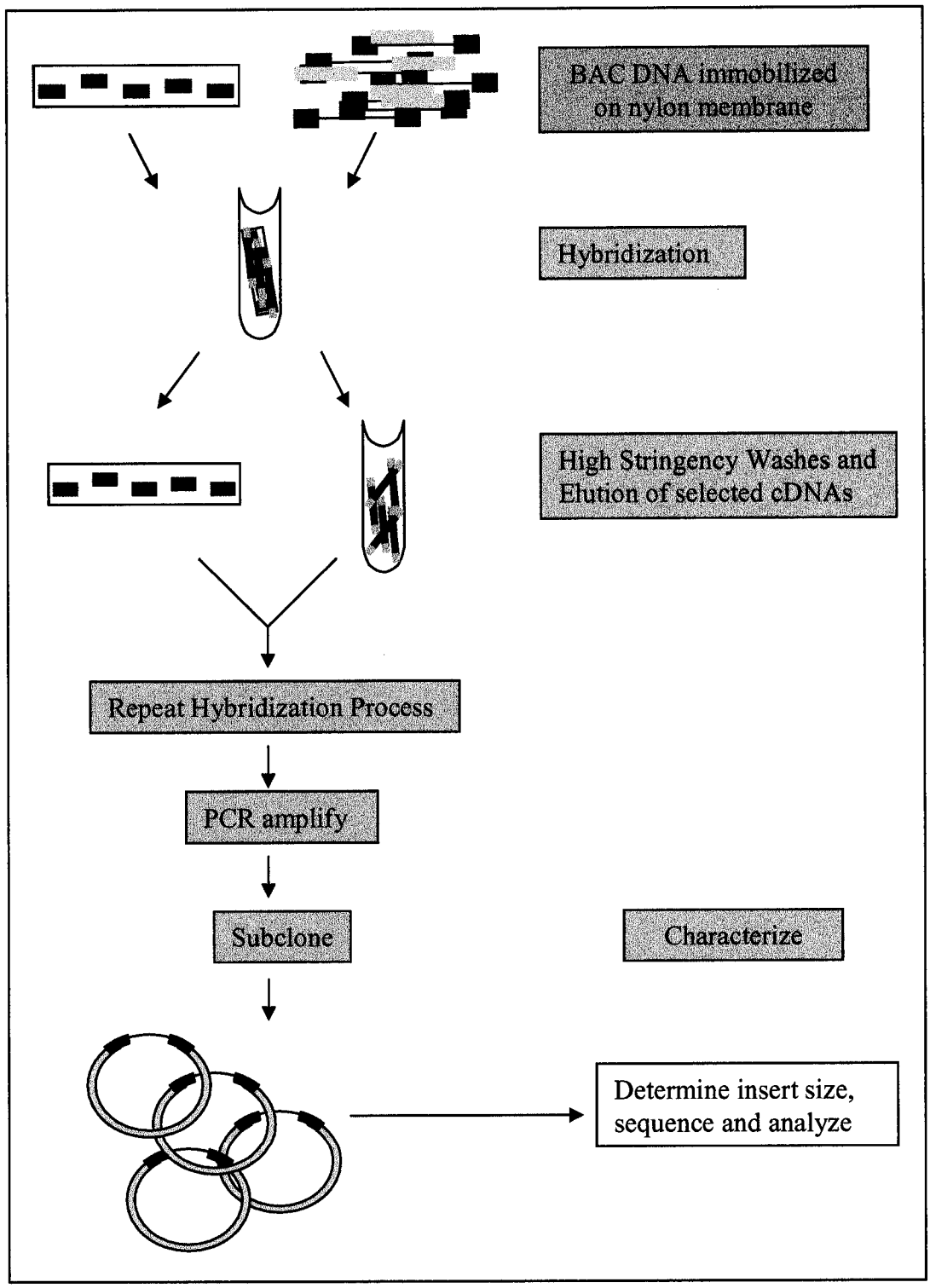
Membranes were prehybridized in hybridization solution (0.5 M sodium phosphate buffer pH 7.2 with 7% SDS, 1mM EDTA and 1 % BSA) at 60°C for 30 min. The preannealed cDNA, the prehybridized membranes containing the BAC DNA and 700 – 800 μl fresh, pre-warmed hybridization solution were combined in a 1.5 mL microcentrifuge tube such that the concentration of cDNA was approximately 12 μg/mL with 1.5 cm^2 of membrane per mL. The membrane and hybridization solution was

incubated at 60°C for 3 days. A generalized overview of the direct selection protocol is outlined in Figure 2.2.

2.3.2 Elution and Second Hybridization of cDNAs

Membranes were removed from the hybridization mixture and washed 3 times with wash buffer 1 (2XSSC, 0.1% SDS) at 20°C for 5, 10 and 20 min durations. Three additional washes with wash buffer 2 (0.2XSSC, 0.1% SDS) at 60°C for 20 min durations were followed by two rinses in 2XSSC and the membranes were placed in a clean 1.5 mL microcentrifuge tube. Sterile, distilled water (250 µl, or enough to submerge the membranes) was added to the tube and it was heated to 100°C for 10 min. 200 µl of the ddH_2O was removed and replaced with 200 µl of fresh ddH_2O and the membranes were heated to 100°C for a further 10 min. 200 µl from each elution was pooled and the membranes were stored at 4°C with the remaining ddH_2O. 20 µg of human placental DNA, 3 M NaOAc, pH 7.2 to a final concentration of 0.3 M and 2.5 volumes of ethanol were added to the collected cDNA and precipitation was carried out over night (O/N) at -20°C. The DNA was pelleted by centrifugation at 14000 rpm for 12 min. and the pellet was resuspended in 0.2 M NaOH and 1 mM EDTA and placed at 60°C for 10 min. for denaturation. The pH was neutralized with 1/10 volumes of 2 M NH_4OAc, pH 5.4 and ¼ volume of 20XSSC was added. Preannealing and the second round of hybridization was carried out in the same manner as the first hybridization.

Figure 2.2. A schematic overview of the cDNA selection protocol. The BACs are immobilized on a nylon membrane followed by hybridization to cDNAs prepared as outlined in Figure 2.1. Following high stringency washing and elution in ddH_2O at 100°C, selected cDNAs are hybridized to the same membrane a second time and eluted in the same manner in order to optimize the selection. Sub-cloning into the *EcoR1* site of pBluescript is followed by analysis of insert size and bi-directional sequencing of clones with inserts greater than 100 bp. Further characterization involves extensive database mining.


BAC DNA immobilized on nylon membrane
Hybridization
High Stringency Washes and Elution of selected cDNAs
Repeat Hybridization Process
PCR amplify
Subclone
Characterize
Determine insert size, sequence and analyze

2.3.3 Final Elution and Amplification of selected cDNAs

The membranes were washed and the cDNAs eluted in the same manner as described for the first hybridization. The selected cDNAs were applied to a Sephadex G25 spin column (Pharmacia) for purification and removal of primers, free nucleotides and small fragments. Selected cDNA was stored frozen at -20°C in the remaining ddH_2O.

Analytical PCR was carried out with small aliquots of the final elution volume. The standard PCR conditions were 94°C for 45 sec for denaturation, annealing at 55°C for 45 sec and extension at 72°C for 2.5 min, for a total of 32 cycles. The RXG oligonucleotide was used as the amplification primer, this ensures amplification of only cDNA fragments that were selected from the cDNA pool. Several individual reactions were pooled and purified using GFX columns (Amersham) to concentrate the cDNA.

Pooled PCR products were digested with *EcoR1* (5 units, 37°C for 2 hours) and purified from the reaction mixture with GFX columns. The vector pBluescript (Stratagene) was digested with *EcoR1* in a similar reaction, purified with GFX columns, dephosphorylated with Shrimp Alkaline Phosphatase (Roche Diagnostics Scientific, Toronto ON) by incubating at 37°C for 30 min. followed by inactivation of the enzyme for 20 min. at 70°C. After purification with a GFX column, ligation reactions were preformed with the Rapid Ligation Kit (Boehringer Mannheim) and transformations were carried out with XL1-blue competent cells prepared as described [83] and 2 μl aliquots of the ligation reaction mixtures according to standard protocols.

2.4 Clone Selection and Characterization

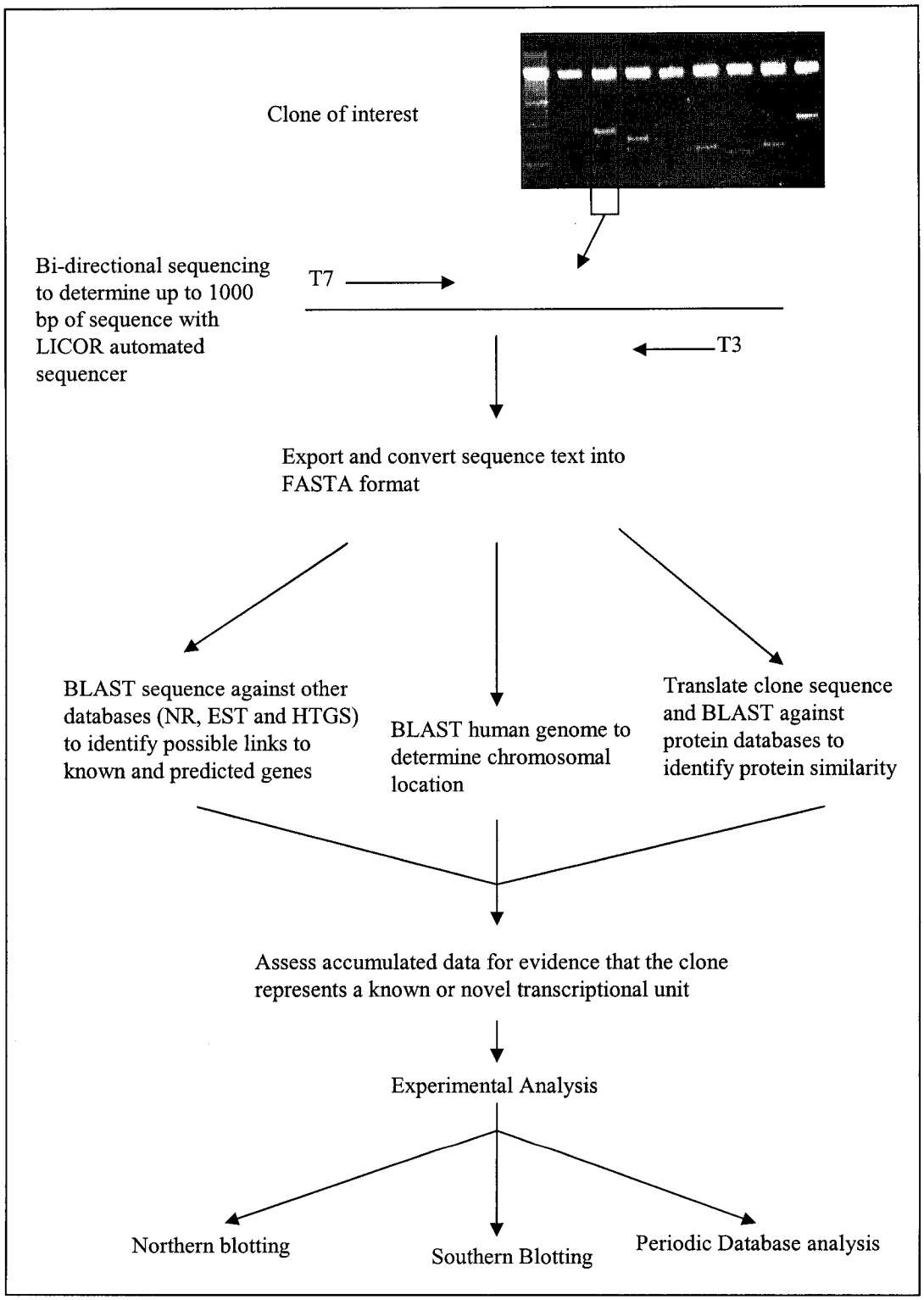
2.4.1 Colony selection and DNA isolation

Single colonies from the transformation were picked with sterilized tooth picks, gridded onto plates and used to inoculate 3-5 mL LB-broth in the presence of ampicillin, 100 μg/ml) for subsequent isolation of plasmid DNA using minipreps (QIAGEN, Mississauga ON). Colony lifts for repeat screening were performed by placing circular nylon membranes directly on the gridded plates as described [81]. Plasmid DNA was digested with *EcoR1* to release the insert, and then subjected to electrophoresis to separate by size on a 1.5% agarose gel.

2.4.2 Clone analysis by Automated Fluorescence Sequencing

An overview of the characterization procedure utilized for each clone is outlined in Figure 2.3. Following determination of clone size, inserts greater than 100 bp were sequenced using the DYEnamic Direct Cycle Sequencing kit (Amersham) using the sequencing sites (T7 and T3) in the vector flanking the insert.. T3-IRD800 and T7-IRD700 primers (Li-Cor, Lincoln Nebraska) were used such that bi-directional sequencing reactions produced sequence from both 5' and 3' regions of each clone in one reaction. Ideally, overlap between the sequences confirmed that the entire insert sequence was obtained in each reaction. Denaturing polyacrylamide gel electrophoresis was carried out with 5-6 % acrylamide (3 mL 10 X TBE, 3.6 mL SequaGel XR [National Diagnostics, Atlanta Georgia USA], 23.4 mL 54% Urea, 150 μl Ammonium persulphate and 15 μl TeMed) and samples were electrophoresed for 9 hours at a constant 2000 V on

Figure 2.3. An overview of the typical clone analysis procedure. Initial bi-directional sequencing and analysis of clones utilizing Repeatmasker identified repetitive clones that were not further characterized. Clones that were found to contain little or no repetitive sequence were analysed with extensive database mining to identify chromosomal location and similarities to known proteins, ESTs or mRNAs. Clones of interest were further analysed with experimental procedures including Northern and Southern blotting. Databases were searched regularly to address the constant evolution of the information within the databases.


Clone of interest
Bi-directional sequencing to determine up to 1000 bp of sequence with LICOR automated sequencer
T7
T3
Export and convert sequence text into FASTA format
BLAST sequence against other databases (NR, EST and HTGS) to identify possible links to known and predicted genes
BLAST human genome to determine chromosomal location
Translate clone sequence and BLAST against protein databases to identify protein similarity
Assess accumulated data for evidence that the clone represents a known or novel transcriptional unit
Experimental Analysis
Northern blotting
Southern Blotting
Periodic Database analysis

a LI-COR DNA Sequencer model 4000, and analysed with the eSeq DNA Sequencing and Analysis software (Li-Cor).

2.4.3 Initial *In Silico* analysis

Sequences from each clone were transformed into FASTA format and subjected to alignment searches using the BLAST software [53]. The NCBI provides data analysis and retrieval resources that operate on the data within a variety of databases accessible through their website (www.ncbi.nlm.nih.gov). Data retrieval resources available through NCBI include Entrez, Pubmed and GenBank [37,54]. Data analysis resources include BLAST, OrfFinder, Unigene, Davis Human-Mouse Homology map, and Online Mendelian Inheritance in Man (OMIM) to name only a select few. A comprehensive outline of these programs was made available by NCBI [84,85,86,87,88,89,90]. An in-depth outline of the processes followed for effective usage of the databases is given in a recent publication in Nature Genetics [91]. Clones were grouped into several categories; 1. Clones that map directly into the 4q34 region and consist of multiple exons, 2. Other clones mapping back to the region, 3. Clones that map to other areas of the genome with high similarity, 4. Clones that do not map to the human genome. Emphasis was placed on clones that mapped to the 4q34 region.

2.4.4 Multiple Tissue Northern Blots

For transcript analysis, two "*Human PolyA$^+$ RNA blots*" were purchased from OriGene Technologies, Inc. (Rockville, MD). Each blot contains 12 major adult tissues

including brain, colon, heart, kidney, liver, lung, muscle, placenta, small intestine, spleen, stomach and testis. Both blots were from the same lot (A1007, Catalogue # HB-2010) and were stored, probed and stripped using the manufacturer's protocols. ULTRAhyb™ ultrasensitive Hybridization buffer (Ambion, Austin TX) was used for hybridization as modified by OriGene Technologies. Actin was provided by OriGene for evaluation of RNA loading on the blots as per the manufacturer's instructions.

2.4.5 Preparation of Radio-labelled Probe

Clones were labelled using [α-^{32}P]dCTP isotope (Amersham, 3000 Ci/mmol) to facilitate hybridization to Southern and Northern blots for analysis. Two techniques were employed. *Random Primed DNA Labelling*: Following *EcoR1* digestion of an individual clone, the cDNA sequence was separated from the vector DNA by electrophoresis on a 1 % agarose gel and extracted from the gel (Gel extraction kit, Qiagen). The Random primed labelling kit (Boehringer Mannheim) was utilized following the manufacturers protocol to radioactively label each clone. Alternatively, *PCR labelling* was undertaken if appropriate primers were available. A standard PCR was set-up with reduced dCTP and the addition of 5 µl of labelled [α-^{32}P]dCTP (10 µC1/µL). Unincorporated nucleotides were removed by application of samples to a G-25 Micro-spin™ column (Amersham). Following either type of radioactive labelling, a 2 µL aliquot of each sample was diluted in 10 mL of Safe-Scint 2® (Intersciences Inc. Markham ON) and a Beckman LS 6500 multipurpose scintillation counter was used to estimate the cpm/µl for each sample. For hybridization to Northern and Southern blots, 3 X 10^6 cpm/mL hybridization solution was used.

2.4.6 Washing and Exposure of Southern and Northern Blots

Washes were undertaken depending on the hybridization of each individual probe. Standard wash buffers (Buffer 1 – 2 X SSC, 0.1 % SDS, Buffer 2 – 0.5 X SSC, 0.1 % SDS, Buffer 3 – 0.2 X SSC, 0.1 % SDS) were used successively (1 – 3) to wash each blot. Between 15-minute washes, the blots were screened with a Geiger counter (Ludlum Measurements Inc., Sweetwater TX) to assess background and specific signals. Following washing, blots were sealed with plastic wrap and used to expose KODAK Biomax MS film. Exposure times were clone dependent and varied from several hours to several days for Southern blots and were generally about 1 week for Northern blots with the exception of LOC133123 which was sufficiently exposed in ~4 hours.

2.5 Further Characterization

2.5.1 Radioactive Sequencing for mutational analysis

Mutation analysis of candidate genes in the 4q32-q34 region was undertaken utilizing radioactive DNA sequencing, patient DNA samples and oligonucleotides designed to amplify individual exons of candidate genes. Each exon was amplified in a standard PCR reaction from PD patient and control DNA. Treatment with Shrimp Alkaline Phosphatase (SAP) and Exonuclease 1 (Exo) was undertaken (1U SAP 1U Exo, 37°C for 20 min., 70°C for 10 min.) to removed unincorporated nucleotides and primers.

Each isolated fragment from above was sequenced using four reactions each containing one [α-^{33}P] labelled nucleotide (G,A,T or C) and the sequencing reactions were electrophoresed on a 5 % polyacrylamide gel (National Diagnostics) at 100 W for

approximately 2 hours. The gel was dried under vacuum for 1 hour and allowed to expose radiographic BIOMAX™ MS (KODAK Scientific Imaging Systems Rochester NY) or AGFA ORTHO HT-G (AGFA Corporation USA) film at -80°C for 1–10 days depending on signal intensity. The exposed film was developed and the results were analysed by manually reading the sequence followed by conversion to FASTA format, which is required for analysis with BLAST software.

Chapter 3

3.0 Results

3.1 Choosing a region for transcription mapping

With the advent of the sequencing of the human genome comes the need to identify all of the transcriptional units within it. We have chosen a region at 4q32-q34 in which we have mapped a novel locus for Parkinson's disease (PD). My work was directed at establishing a transcription map of a portion of the PD region, for which the results will be used in the identification of a PD gene. The candidate locus, 4q32-q34, encompasses approximately a 20 Mb region. A BAC physical map had been constructed in our lab based on the information available in various databases and BAC fingerprinting.

3.2 The 4q32 region

Due to the large size of the PD candidate region (20 Mb), only a portion of the region would be chosen to analyse with direct selection. The region chosen for the initiation of transcription mapping was 4q32.

3.2.1 The physical map of 4q32

The BAC physical map of the 4q32 region contained 32 BACs in both finished and draft quality. Analysis of the composition of the BACs in this region is outlined in

Table 3.1. These BACs made up approximately 6 Mb of sequence spanning 4q32.2 – q32.3 in May 2001. An illustration of the map at the time of the initiation of this research is shown in Figure 3.1. At that time, the region contained 6 gaps of unknown size. As a result, the actual size of the region included in the physical map could not be accurately predicted. Therefore, it was understood that the size and orientation of the map could drastically change during the course of our work. Currently, the map spans one genomic contig, NT_006171, and represents approximately 4.5 Mb of sequence. There are no remaining gaps and only two BACs remain in draft sequence quality. The gaps that were present in the original map have been filled with BACs that were not included in my analysis.

3.2.2 Genes in the 4q32 region

The entire 6 Mb of sequence in the 4q32 region contained only two known genes, neuropeptide Y receptors 1 and 5 (*NPYR1* and *NPYR5*, respectively). These transcripts are both located in the BAC clone 719L21, organized in an overlapping structure indicating they may have resulted from a gene duplication and may be transcriptionally regulated in coordination [92]. Neuropeptide Y receptor proteins are postulated to be involved in the regulation of feeding behaviour and are primarily expressed in the central nervous system [93].

The region also contained 10 predicted genes based on EST, cDNA and mRNA evidence. Seven of the 10 predicted transcripts are encompassed within approximately 1 Mb of sequence in the region surrounding the *NPYR* locus. One of the predicted genes that was located outside of the immediate area of the *NPYR* locus, DKFZp566D234, has

Table 3.1. The BACs used for direct selection at 4q32.

BAC Clone (RP11, AC0)	Size (bp)	Genes in region	Sequence State	% Repeat	% G-C
99N09 AC105251	44971	ESTs	complete	36.54	33.78
10N03 AC096717	109027	ESTs, PP	complete	36.77	33.62
497K21 AC023136	185593	ESTs, PP	complete	37.02	33.05
694K14 AC108163	95304	ESTs, PP	complete	35.81	33.85
82A1 AC007733	162737	None	Draft	49.64	34.56
502F14 AC060818	194103	EST, PP	Draft (chromo2)	51.18	34.84
1G08 AC105284	144983	PP	Draft	43.07	34.64
422D16 AC114770	127953	PP	finished	51.25	35.24
277G18 AC096563	110282	PP	finished	52.01	35.95
487M20					
493C20 AC098867	142314	EST, PP	finished	43.81	36.73
326P21 AC092631	191468	EST, PP	finished	42.12	35.45
277J14 AC096563	110282	None	finished	52.01	35.95
57B18					
808H17 AC079240	202847	EST, PP (introns only)	finished	39.29 deleted record	35.22
402 D23 AC021134	135692		finished	40.10	36.15
563E02 AC022272	186535	ESTs, PP	finished	52.53 lots of LTR	38.29
719L21 AC079238	142462	2KG	finished	41.82	37.52
218F10 AC093788	160853	EST, PP	finished	44.40 deleted record	37.16
396H20 *					
39C10 AC105250	70449	EST, PP (introns only)	finished	33.87	35.68
360H22 AC079232	169281	EST, PP (intron only)	finished	52.25	35.13
527F14 AC068548	42137	Pp	finished	53.79	39.04
75A5 AC074198	182178	PP	finished	56.80	38.51
294 N02 AC036112	66863	PP	draft	36.73	41.51
502M1, AC093853	212207			50.06	35.60
219G14, AC095047	72660		Finished	42.99	34.44
48F23, AC115222	127707		Finished	55.82	35.15
23406 AC104793	133244	PP	Finished	45.01	35.19
47G2 AC072037	143484	None	Draft Chromosome 3	52.58	34.77
420J17 AC013328	181278	PP, EST	Draft Chromosome 3	49.37	34.68
20K23 AC016174	161304		Draft	45.66	35.32

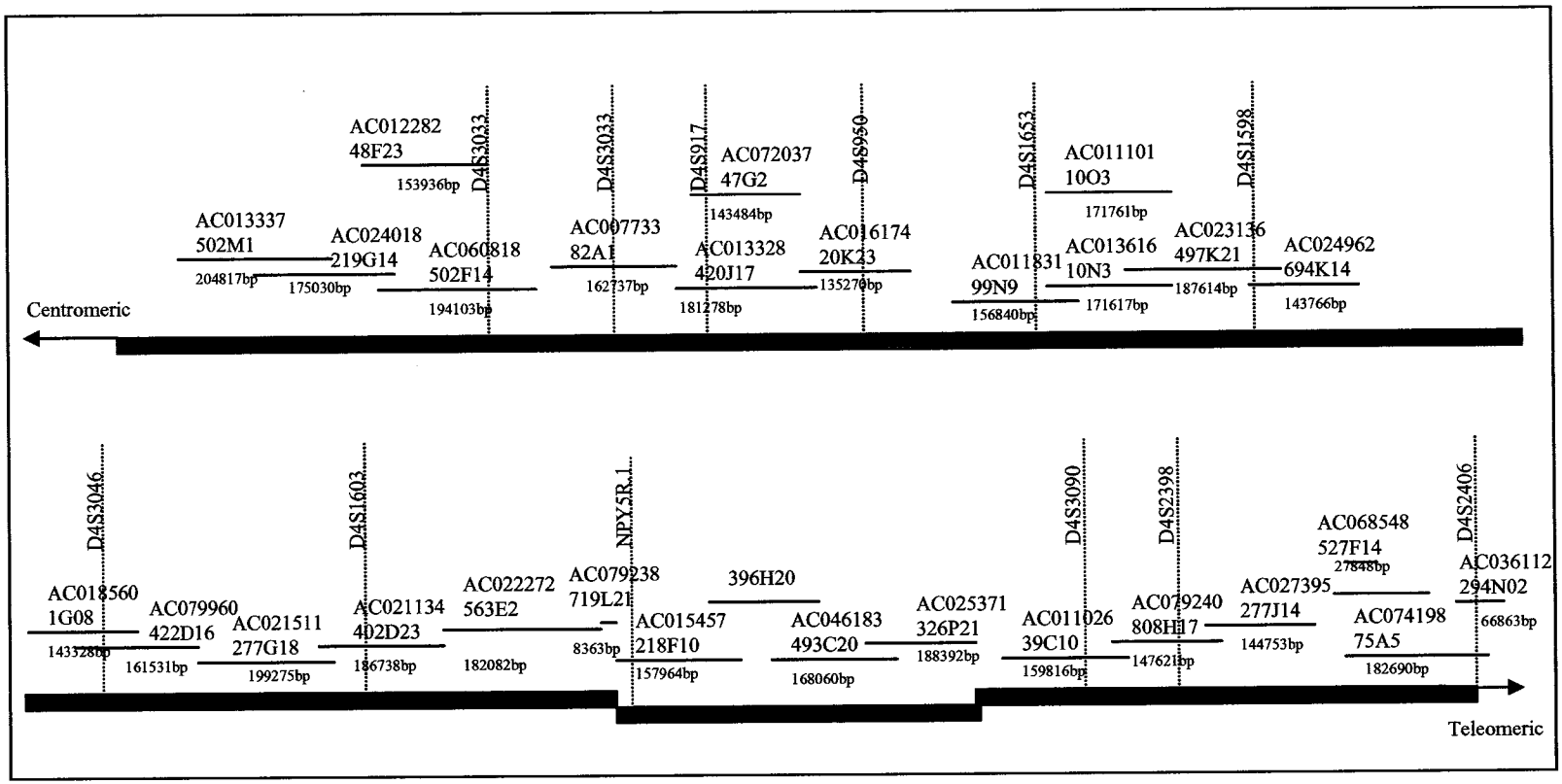
EST - expressed sequence tag, PP – predicted protein, KP – known protein uncharacterized, KG - known gene.

Note - repeat analysis of draft clones is inaccurate due to strings of n's used to separate fragments

- BACs are listed in centromeric to telomeric order based on the Sept. 2001 physical map.

* BACS with no information and are no longer available in the databases.

Figure 3.1. The 4q32 map from May 2001. Information was compiled from the NCBI website. Microsatellite markers (D4S) are shown as dotted vertical lines. BACs are labelled with both accession (AC) and Roswell Park Cancer Institute library designations. Areas where BACs do not overlap are depicted by gaps in the map. All sizes are approximate.


AC012282
48F23
153936bp
D4S3033
D4S3033
D4S917
AC072037
47G2
143484bp
D4S950
D4S1653
AC011101
10O3
171761bp
D4S1598
AC013337
502M1
204817bp
AC024018
219G14
175030bp
AC060818
502F14
194103bp
AC007733
82A1
162737bp
AC013328
420J17
181278bp
AC016174
20K23
135276bp
AC011831
99N9
156840bp
AC013616
10N3
171617bp
AC023136
497K21
187614bp
AC024962
694K14
143766bp
Centromeric
D4S3046
D4S1603
NPY5R.1
D4S3090
D4S2398
D4S2406
AC068548
527F14
27848bp
AC036112
294N02
66863bp
AC018560
1G08
143328bp
AC079960
422D16
161531bp
AC021511
277G18
199275bp
AC021134
402D23
186738bp
AC022272
563E2
182082bp
AC079238
719L21
8363bp
396H20
AC015457
218F10
157964bp
AC046183
493C20
168060bp
AC025371
326P21
188392bp
AC011026
39C10
159816bp
AC079240
808H17
147621bp
AC027395
277J14
144753bp
AC074198
75A5
182690bp
Teleomeric
Genomic Contig
BAC
Genetic marker

supporting EST evidence and spans 625 kb of the 4q32 region. The mRNA for this hypothetical protein is 1.72 kb, based on 6 supporting clones and is incomplete at the 3' end. The function of this partial protein (438 aa, MW 49.7 kDa, pI 5.8) is unknown [94].

Based on the relatively low gene density of the 6 Mb region at 4q32, and the presence of several large regions with no annotated genes, it was hypothesized that cDNA selection utilizing the BACs in this region would identify novel transcriptional units.

3.3 Direct selection at 4q32

3.3.1 *NPYR3* a positive control

An initial simple direct selection experiment was designed as a positive control experiment. The BAC 487M20 was shown by NCBI to contain the neuropeptide Y receptor protein, *NPYR3*. This BAC was located on chromosome 4, outside of the 4q32 region. *NPYR3* was chosen as the positive control because it was outside of the 4q32 region and the *NPYR3* gene was expressed mainly in brain. If this initial control experiment was successful, *NPYR3* would serve as a positive control for each subsequent direct selection by inclusion of this BAC in each direct selection. Furthermore, this experiment would serve to illustrate the type of result expected from a direct selection, being the first such attempt undertaken in our lab.

We first evaluated the substantia nigra (SN) cDNA pool generated in our lab for the presence of *NPYR3* by hybridizing the SN cDNA pool to the BAC 487M20 that had been immobilized on nylon membrane by dot blotting. Following direct selection, the

isolated cDNAs were amplified by PCR and sub-cloned into pBluescript. Of 48 clones analysed, 28 contained inserts, and the average insert size was 800 bp. Inserts were sequenced and analysed for repetitive elements utilizing the Repeatmasker webserver, which did not identify any common repetitive elements like LINEs or SINEs. However, analysis with the BLAST software and the various human databases revealed that all clones were highly represented elements from the human genome including mitochondrial sequences, dihydrofolate reductase that is represented on several chromosomes, an oxygenase reductase that is also highly represented in the genome and rRNA. No clones were identified that contained any portion of the *NPYR3* transcript.

A detailed analysis of the BAC 487M20 utilizing the information in the various databases indicated that the presence of *NPYR3* in the BAC 487M20 was ambiguous because *NPYR3* appeared to be localized to more than one chromosomal region. Attempts to amplify sequence from *NPYR3* with PCR from the BAC 487M20 and from our cDNA pools were negative (data not shown). Following these experiments, the *NPYR3* gene was removed from our region of chromosome 4 according to the NCBI database, and 487M20 was removed from the database altogether. Currently, Jan. 2003, there is no *NPYR3* annotated in the human genome.

3.3.2 Direct selection in a sub-region of 4q32

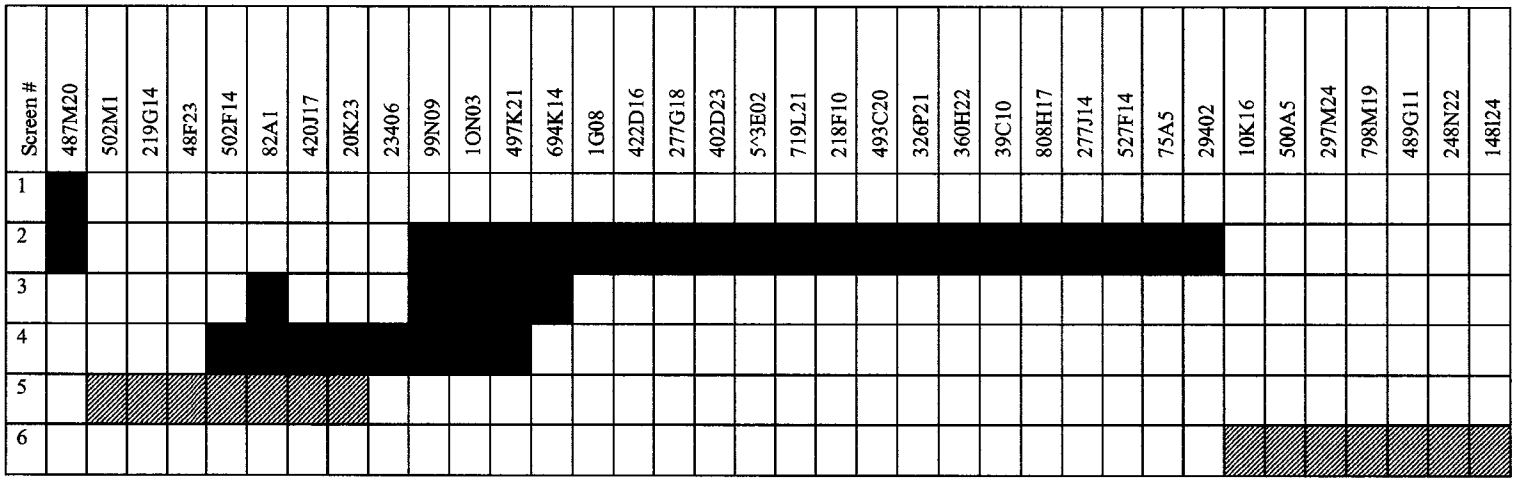
To identify novel transcriptional units in the 4q32 region, the next experiment was designed to undertake direct selection using BACs spanning the entire 6 Mb 4q32.2 – q32.3 region. It was postulated that successful direct selection would identify transcriptional units representing both known and predicted genes, which would validate

the technique with respect to our region, and act as a positive control. Concurrently we would identify unknown transcriptional units, which could be further characterized.

The BACs were divided into "pools" based on their overlap on the physical map. Four pools of 5 BACs and one pool of 6 BACs contained all of the BACs that were available in our lab covering the 4q32 region (Table 3.2). A single dot blot was performed and each pool of BACs was immobilized on a single "dot". The cDNA selection was undertaken in one tube containing all 5 pools, however the final elution was done individually for each pool.

Analysis of the cDNAs isolated with direct selection in the 4q32.2-q32.3 region with sub-cloning and bi-directional sequencing indicated similar results to the positive control experiment summarized in section 3.3.1. The majority of clones represented mitochondrial sequence or rRNA, which were hypothesised to represent a level of background. Several attempts were made to reduce the level of background obtained with direct selection in the 4q32 region. Because the sonicated human placental DNA (competitor DNA) was utilized in a large excess, it was assumed that the background was not due to incomplete blocking. Direct selections were undertaken with fewer BACs, with different cDNA pools and using both dot blots and Southern blots (Table 3.2). All of these attempts gave similar results; clones appeared to be background sequences including repetitive elements such as LINEs and SINEs and highly represented elements containing mitochondrial DNA sequences, and rRNA. However, following sequencing of approximately 300 clones, there were 3 individual clones that were mapped *in silico* and 2 that were mapped by Southern hybridization to the 4q32 region. This indicated that this direct selection may have successfully identified transcriptional units.

Table 3.2. Outline of the BACs used for each round of cDNA selection.



Note - BACs are listed in centromeric to telomeric order based on the physical map from May 2001.

- 487M20 was a positive control outside of the 4q32 region.

Immobilization of BACs

■ Northern blot □ Southern blot

Complementary DNA pool

▧ Substantia nigra ▨ Adult brain, fetal brain, testis, placenta

Clones 1, 41, and 383 were shown to map into the 4q32 region based on *in silico* analysis, and represent intronic sequence of the predicted gene DKFZp566D234 (Figure 3.2). Clones 5, and 41 were shown to map onto the originating BACs by probing a Southern blot of the BACs (Figure 3.3). Sequences for these clones are outlined in Appendix 2. Further *in silico* analysis revealed that all 4 of these clone consisted of greater than 20 % repetitive sequence (Table 3.3 and Appendix 2).

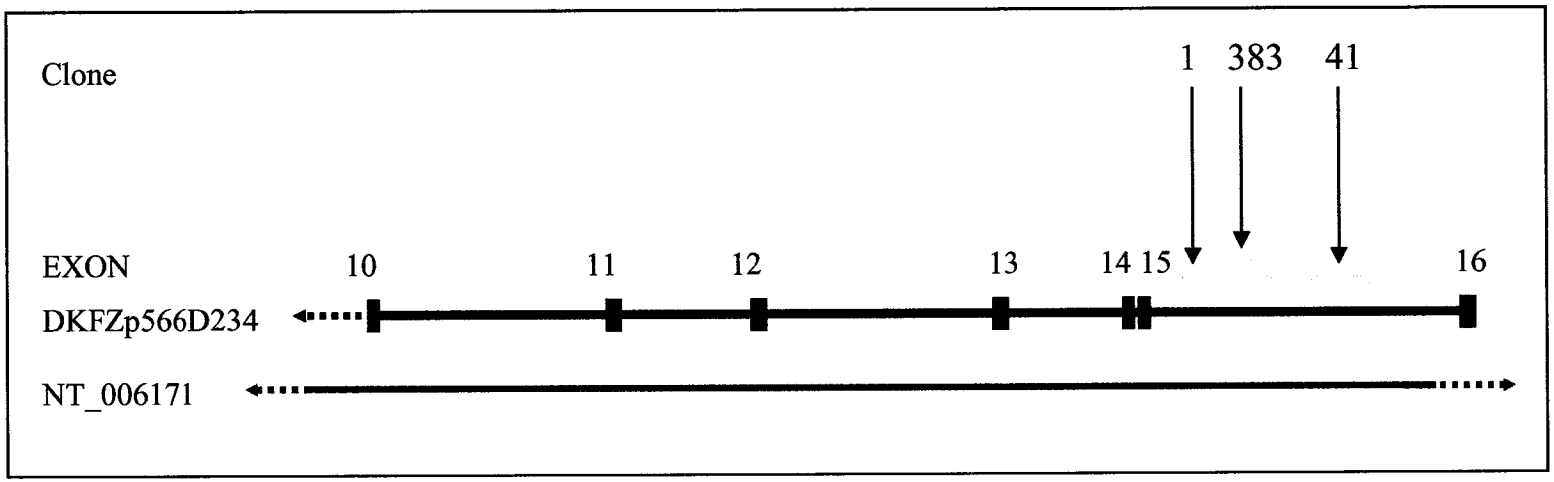
3.3.3 Summary – 4q32

During the course of direct selection in the 4q32 region, more than 600 clones had been isolated and analysed and only 5 clones could be mapped back to the original BAC contig. The physical map in this 6 Mb region represented in the NCBI database was very gene poor. Two known genes, 10 predicted genes, and very few ESTs were present in this region. In order to facilitate a more productive direct selection attempt, it became apparent that a new region would need to be identified from within our PD locus.

3.4 Choosing a new region

In the third attempt to establish a transcript map, a region within the PD candidate region was chosen that had several advantages over the previous regions. A known gene that was placed in the BACs and that was anchored to the genomic region was essential. Next, a region that contained ESTs with no assigned genes would give evidence that there are novel genes to be discovered in the region. Finally, a smaller, more workable number of BACs, in varying degrees of completeness would be chosen to represent the physical map of the next region.

Figure 3.2. Overlap of three cDNA clones (1, 41, and 383) with predicted protein by *in silico* mapping to BAC 694K14. Clone sizes are not to scale. These were the only three clones to be mapped *in silico* into the 4q32 region in this screen. These clones overlap with intronic sequence of the predicted gene DKFZp566D234. Only the 3` portion of this gene, which spans 625 kb, is shown.


Clone
1
383
41
EXON
10
11
12
13
14 15
16
DKFZp566D234
NT_006171

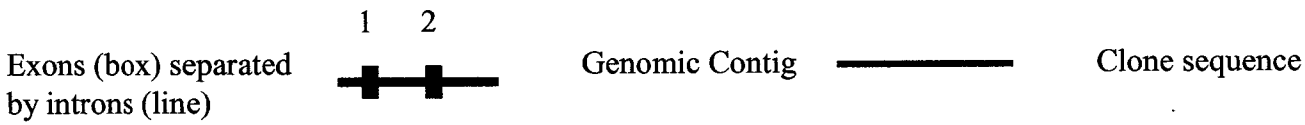

1 2
Exons (box) separated by introns (line)
Genomic Contig
Clone sequence

Figure 3.3. Mapping back of cDNA clones 5 and 41 to BAC DNA by Southern blotting. *EcoR1* digested BACs were hybridized with radioactively labeled cDNAs 5, and 41 shown in A. and B. respectively. The 4q32 BACs are loaded in order from centromeric to telomeric (left to right). **A.** Clone 5 is shown to be present in BAC 82A1. *In silico* analysis demonstrates that it is located on chromosome 2. **B.** Clone 41 is shown to map to BACs 82A1 and 694K14 where it also maps *in silico*. These BACs do not overlap in the current physical map.

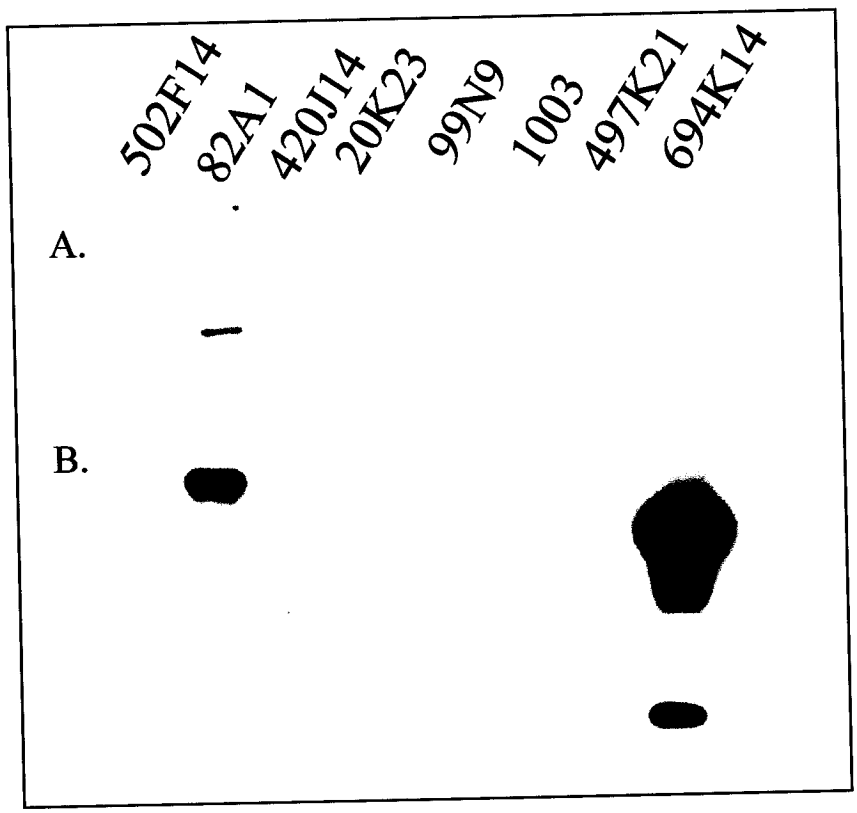

502F14
82A1
420J14
20K23
99N9
1003
497K21
694K14
A.
B.

Table 3.3. Repetitive elements found in the 4q32 clones.

Clone	Repeat	% repetitive
1	LINE	58 %
5	MER 1	30 %
41	MIR 3 (SINE)	21 %
383	MIR 3 (SINE)	21 %

3.5 Direct selection at 4q34

It was found that a region at 4q34 fulfilled all of these requirements. This region spanned approximately 6 cM with a tiling path of 7 BACs (Table 3.4). The BACs were arranged in two contigs with one gap dividing them into a group of 5 and a group of 2. All of the BACs except one (248N22) had known genes strongly anchored on them, with a total of 7 known genes in the region. This genomic region is illustrated in Figures 3.4 and 3.5. All BACs had a number of ESTs and *in silico* predicted genes. Furthermore, mRNA evidence suggested the existence of several transcriptional units within the BACs that were not associated with known genes [94].

3.5.1 Genes in the 4q34 region

One of the strengths of the 4q34 region as a candidate for direct selection was that it contained 7 known genes. These genes would provide a means to identify known transcripts in the 4q34 region with direct selection, thereby providing 7 possible positive controls. Furthermore, the construction of a fine-detailed transcript map of the 4q34 region included an assessment of the previous characterization of the known and predicted transcripts in the 4q34 region (Tables 3.5 and 3.6). A brief overview of each known gene and the *in silico* predicted genes is given in the following section. Southern blotting of each of the originating BACs with the known and predicted genes in the 4q34 region confirmed and identified the placement of the genes within the physical map. Northern hybridizations supplemented the known expression data or confirmed expression of uncharacterized transcripts by analysing the expression of the transcripts in

Table 3.4. Summary of the BACs in the chosen 4q34 region.

Clone	Accession	GI	Status (Sept. 2002)	Bases	% A/T	% repeat	Chromosomal Location
RP11-10K16	AC105285 AC011223	gi:19071675	Finished	156635	60.9 %		179099702 : 179254337
RP11-500A5	AC098596	gi:16418268	Draft	175306	60.6 %		
RP11-297M24	AC009588	gi:10045234	Draft	166472	60.5 %		
RP11-798M19	AC097534 AC022744	gi:16874920	Finished	139455	61.5 %		179258316 : 179397771
RP11-489G11	AC093849 AC015531	gi:16647565	Finished	156195	57.9 %		179397771 : 179551966
RP11-471J12	AC113154	gi:20429610	Finished	42983	61.0 %		179551966 : 179588361
RP11-475B2	AC079789	gi:19698730	Finished	198586	62.0 %		179636209 : 179791696
RP11-717B15	AC109640	gi:18543163	Draft	167512	63.7 %		179791696 : 179874325
RP11-161D15	AC106895	gi:18308813	Draft	179667	63.2 %	35.37 %	179911031 :180021827
RP11-248N22	AC012055	gi:8671944	Finished	171943	63.7 %		180021827 : 180193770
RP11-148L24	AC097653 AC034117	gi:18042366	Finished	145496	64.4 %	32.85 %	180193770 : 180337266
						Approximate region length	1237564 bp
						Gaps in the region	2

Note: BACs are listed in order from most centromeric to telomeric, as represented on NCBI in November 2002.
Shaded boxes indicate BACs that moved into the region after cDNA selection, and were therefore not included in the screen.

Figure 3.4. A detailed illustration of the centromeric region of the 4q34 map. **A.** Clones isolated with the cDNA selection protocol are illustrated in orange. The exon number is designated if clones contained more than one exon. **B.** A portion of the genomic contig NT_006257 is shown and numbered in base pairs. The single gap in this portion of the contig is designated as a dotted line, which divides the map into two pages. **C**. Genes are color-coded depending on their level of supporting evidence; a legend appears on the second page. Genes appear in two rows, the top row represents those oriented on the genomic contig in the sense direction and the bottom row represents those oriented in the antisense direction. Exons are illustrated as vertical lines connected by introns (horizontal line). Genes are shown approximately to scale, although small exons may be compressed and exon sizes are not representative. **D**. BACs are labeled with the designation given by Roswell Park Cancer Institute during library construction, and are illustrated approximately to scale.

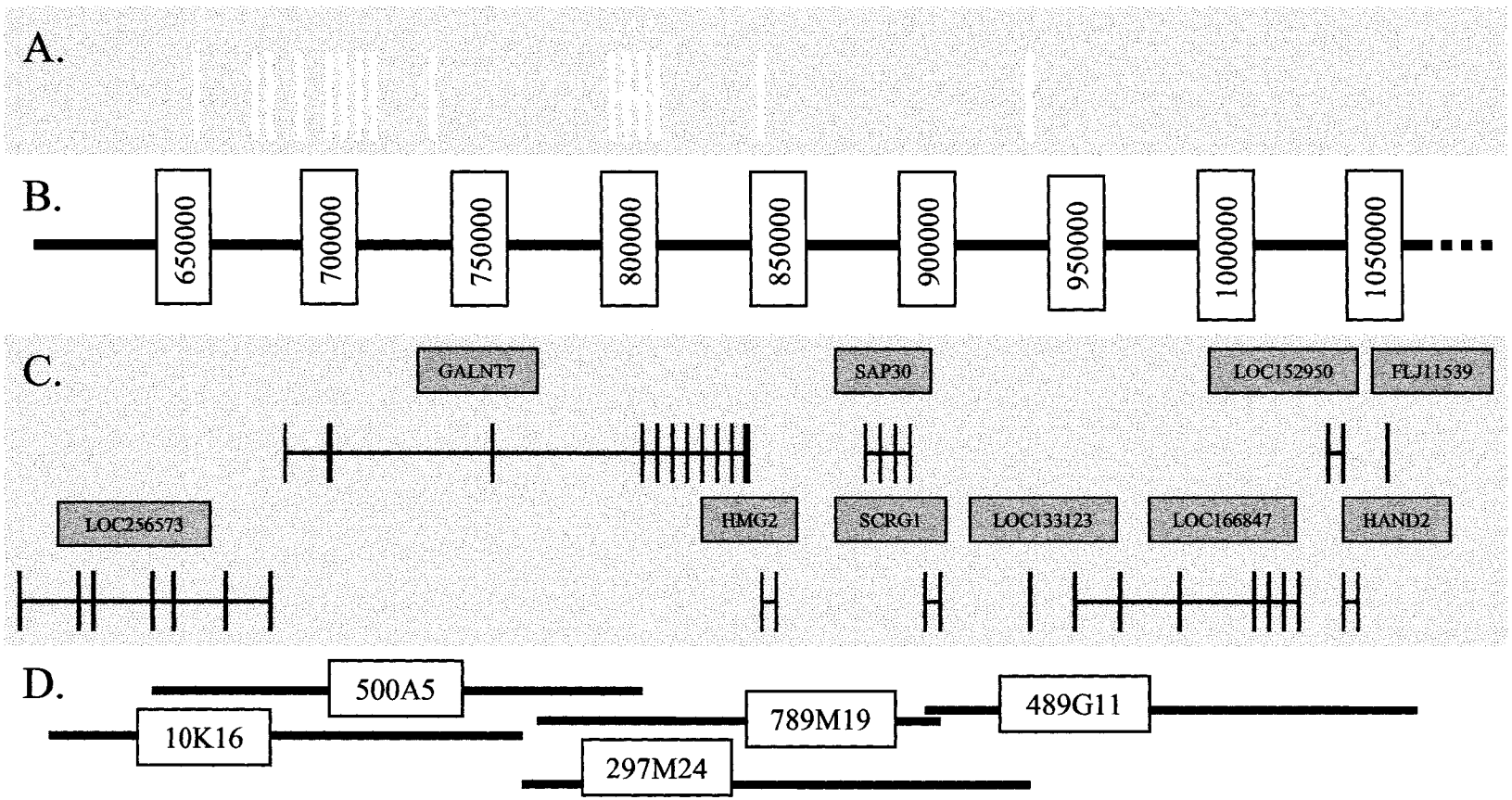

A.
B.
650000
700000
750000
800000
850000
900000
950000
1000000
1050000
C.
GALNT7
SAP30
LOC152950
FLJ11539
LOC256573
HMG2
SCRG1
LOC133123
LOC166847
HAND2
D.
500A5
10K16
789M19
489G11
297M24

Figure 3.5. The telomeric portion of the 4q34 map. Refer to Figure 3.4 for detailed analysis.

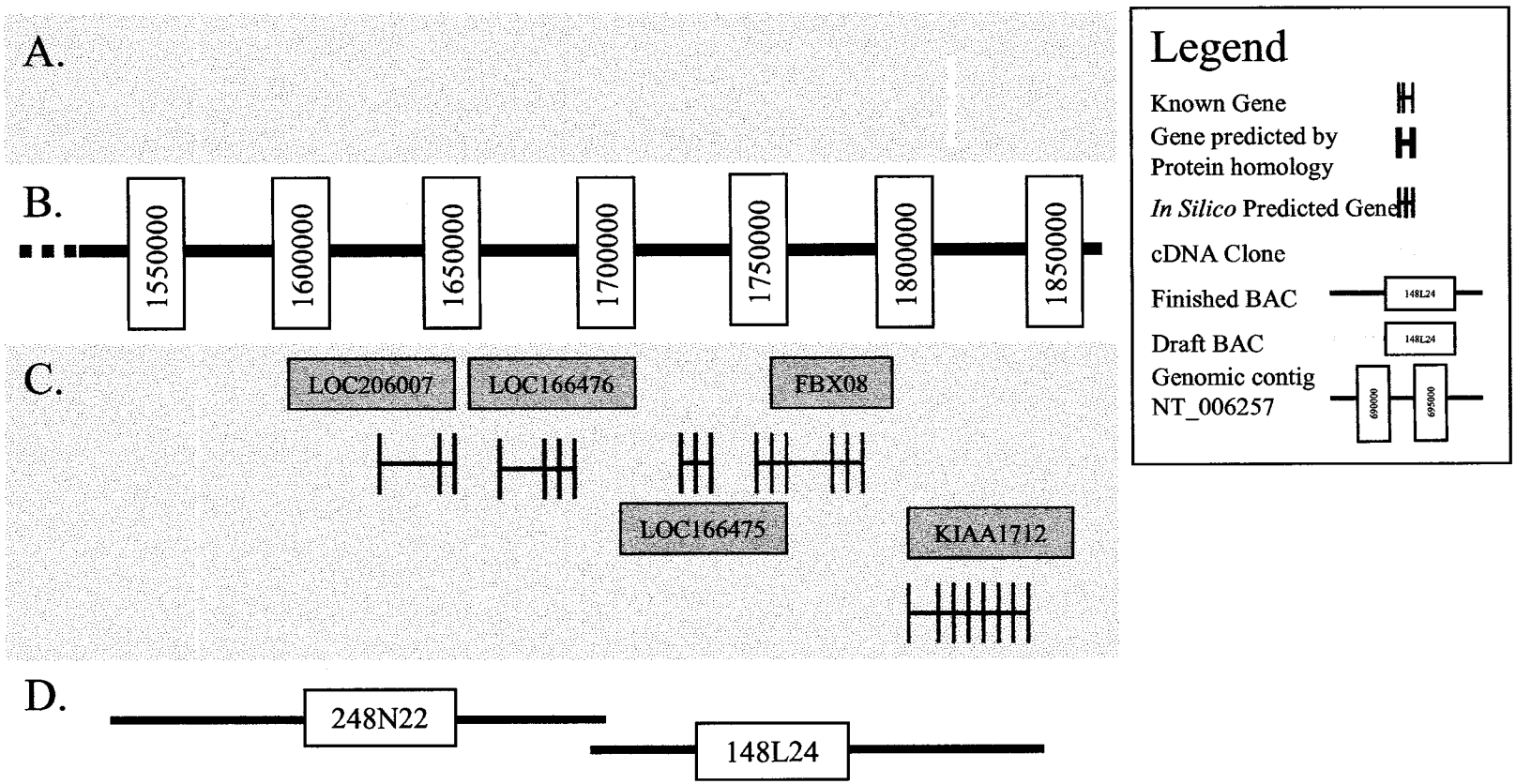

A.
B.
1550000
1600000
1650000
1700000
1750000
1800000
1850000
C.
LOC206007
LOC166476
FBX08
LOC166475
KIAA1712
D.
248N22
148L24
Legend
Known Gene
Gene predicted by Protein homology
In Silico Predicted Gene
cDNA Clone
Finished BAC
148L24
Draft BAC
148L24
Genomic contig NT_006257
600000
650000

Table 3.5. Known genes from the BACs used for cDNA selection

Gene	Name	Transcript size including splice variants	Orientation	Mouse homolog Chromosome 8
GALNT7	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase	1.58 kb, 4.27 kb, 4.26 kb (2), 4.32 kb and 1.89 kb	Direct	Galnt7
HMG2	High-mobility group 2	2.20 kb, 2.23 kb, 2.17 kb, 2.28 kb, 2.16 kb, 2.56 kb, 2.00 kb, 1.65 kb, and 1.36 kb.	Reverse	Hmgb2
SAP30	Sin3 associated protein	1.44 kb (2)	Direct	Sap30
SCRG1	Scrapie-responsive protein 1	0.55 kb, 1.12 kb, 1.19 kb, 0.91 kb	Reverse	Scrg1
HAND2	Heart-and-neuralcrest derivatives expressed 2	1.19 kb, 1.20 kb, 1.58 kb, 1.50 kb	Reverse	Hand2
FLJ11539 (LOC1529520)	Hypothetical protein predicted to be in the mitochondria transcription factor???	1.58 kb, 2.96 kb, 3.00 kb, 2.84 kb, 1.59 kb, 2.39 kb, 1.84 kb	Direct	
FBX08	F-box only protein 8	2.0 kb	Reverse	Fbx08
KIAA1712		4.4 kb	Direct	LOC234276

Table 3.6. The predicted genes in the 4q34 region.

Predicted Transcript	Locus Type	Supporting evidence
LOC256573	model, ab initio, with EST support	similar to putative UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase T9
LOC133123	model, ab initio, with EST support	similar to 60S ribosomal protein L5
LOC166847	model, ab initio, with EST support	
LOC152950	model, supported by EST alignments	
FLJ11539	gene with protein product, function unknown	hypothetical protein FLJ11539
LOC206007	model, ab initio, with EST support	
LOC166476	model, ab initio, with EST support	
LOC166475	model, ab initio, with EST support	
Hs4_6414_29_2_2216	Aceview	1 cDNA
Hs4_6414_29_2_2716	Aceview	2 cDNAs
Hs4_6414_29_2_3431	Aceview	1 cDNA
Hs4_6414_29_3_621	Aceview	1 cDNA
Hs4_6414_29_3_618	Aceview	3 cDNA
Hs4_6414_29_3_582	Aceview	9 cDNA
Hs4_6414_29_3_721	Aceview	6 cDNA
Hs4_6414_29_3_1000	Aceview	10 cDNA
Hs4_6414_29_3_1493	Aceview	1 cDNA
Hs4_6414_29_3_1757	Aceview	1 cDNA
Hs4_6414_29_3_1945	Aceview	3 cDNA
Hs4_6414_29_3_2276	Aceview	4 cDNA
Hs4_6414_29_3_2471	Aceview	4 cDNA
Hs4_6414_29_3_2729	Aceview	1 cDNA

brain, colon, heart, kidney, liver, lung, muscle, placenta, small intestine, spleen, stomach and testis.

3.5.2 *GALNT7*

UDP-N-acetyl-alpha-D-galactosamine:polypeptide-N-acetylgalactosaminyl-transferase 7 (*GALNT7)* is identified as a member of the GalNac-transferase family [95]. The *GALNT7* sequence is found to span 150 kb represented in 4 of the BACs in the 4q34 region (Figure 3.4). Analysis by Southern blotting confirmed the location of *GALNT7* in the BACs 10K16, 500A5, 789M19 and 297M24 (Figure 3.6). *GALNT7* consists of 18 exons, 10 of which are alternative. *GALNT7* has six published alternative transcripts (Figure 3.7). The major transcript is 4.26 kb as reported in the public database [54]. This gene functions in the initiation step of O-glycosylation and transfer of N-acetylgalactosamine to serine and threonine residues. It is a type II transmembrane protein and shares sequence motifs with other GALNT family members.

Previous characterization by Bennett *et al.* indicates that *GALNT7* is expressed in various tissues including stomach, thyroid, spinal cord, lymph node, trachea, adrenal gland and bone marrow [96]. Characterization with Northern hybridization indicates expression in brain, kidney, lung, small intestine, and testis with low levels in placenta (Figure 3.8). Verification of the previously reported expression in stomach was not obvious.

brain, colon, heart, kidney, liver, lung, muscle, placenta, small intestine, spleen, stomach and testis.

3.5.2 *GALNT7*

UDP-N-acetyl-alpha-D-galactosamine:polypeptide-N-acetylgalactosaminyl-transferase 7 (*GALNT7)* is identified as a member of the GalNac-transferase family [95]. The *GALNT7* sequence is found to span 150 kb represented in 4 of the BACs in the 4q34 region (Figure 3.4). Analysis by Southern blotting confirmed the location of *GALNT7* in the BACs 10K16, 500A5, 789M19 and 297M24 (Figure 3.6). *GALNT7* consists of 18 exons, 10 of which are alternative. *GALNT7* has six published alternative transcripts (Figure 3.7). The major transcript is 4.26 kb as reported in the public database [54]. This gene functions in the initiation step of O-glycosylation and transfer of N-acetylgalactosamine to serine and threonine residues. It is a type II transmembrane protein and shares sequence motifs with other GALNT family members.

Previous characterization by Bennett *et al.* indicates that *GALNT7* is expressed in various tissues including stomach, thyroid, spinal cord, lymph node, trachea, adrenal gland and bone marrow [96]. Characterization with Northern hybridization indicates expression in brain, kidney, lung, small intestine, and testis with low levels in placenta (Figure 3.8). Verification of the previously reported expression in stomach was not obvious.

Figure 3.6. Southern Blot analysis of known and predicted genes within the BAC tiling path of 4q34. BACs are arranged on the blots in order from centromeric to telomeric (left to right) and were probed with radioactively labelled *GALNT7*, *HMG2*, *FBX08*, KIAA1712, *SCRG1*, LOC152952, LOC133123, LOC166847, LOC152950, and LOC166475 respectively. Genes predicted based on *in silico* analyses are annotated beginning with "LOC". Exposure times varied from 2 hours to 2 days.

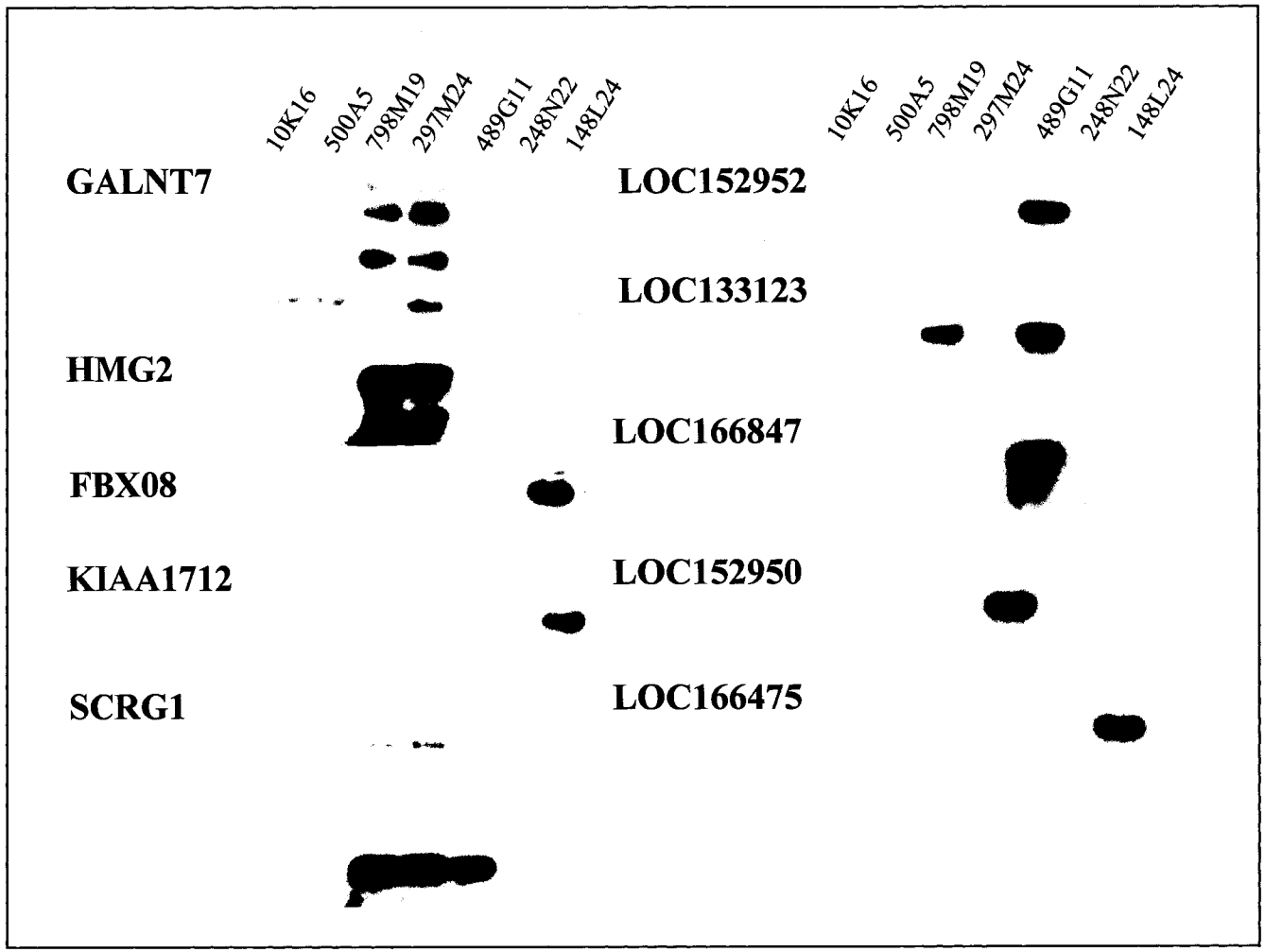

10K16
500A5
798M19
297M24
489G11
248N22
148L24
GALNT7
HMG2
FBX08
KIAA1712
SCRG1
10K16
500A5
798M19
297M24
489G11
248N22
148L24
LOC152952
LOC133123
LOC166847
LOC152950
LOC166475

Figure 3.7. The documented splice variants of *GALNT7*. **1.** There are six alternative transcripts annotated in the Aceview database. Transcripts vary in size from 1.58 kb to 4.32 kb, however, the major transcript is 4.26 kb and is represented as isoform "B". **2.** The cDNA-selected clone 1040 does not overlap with any of the documented transcripts because it does not contain an exon 6.

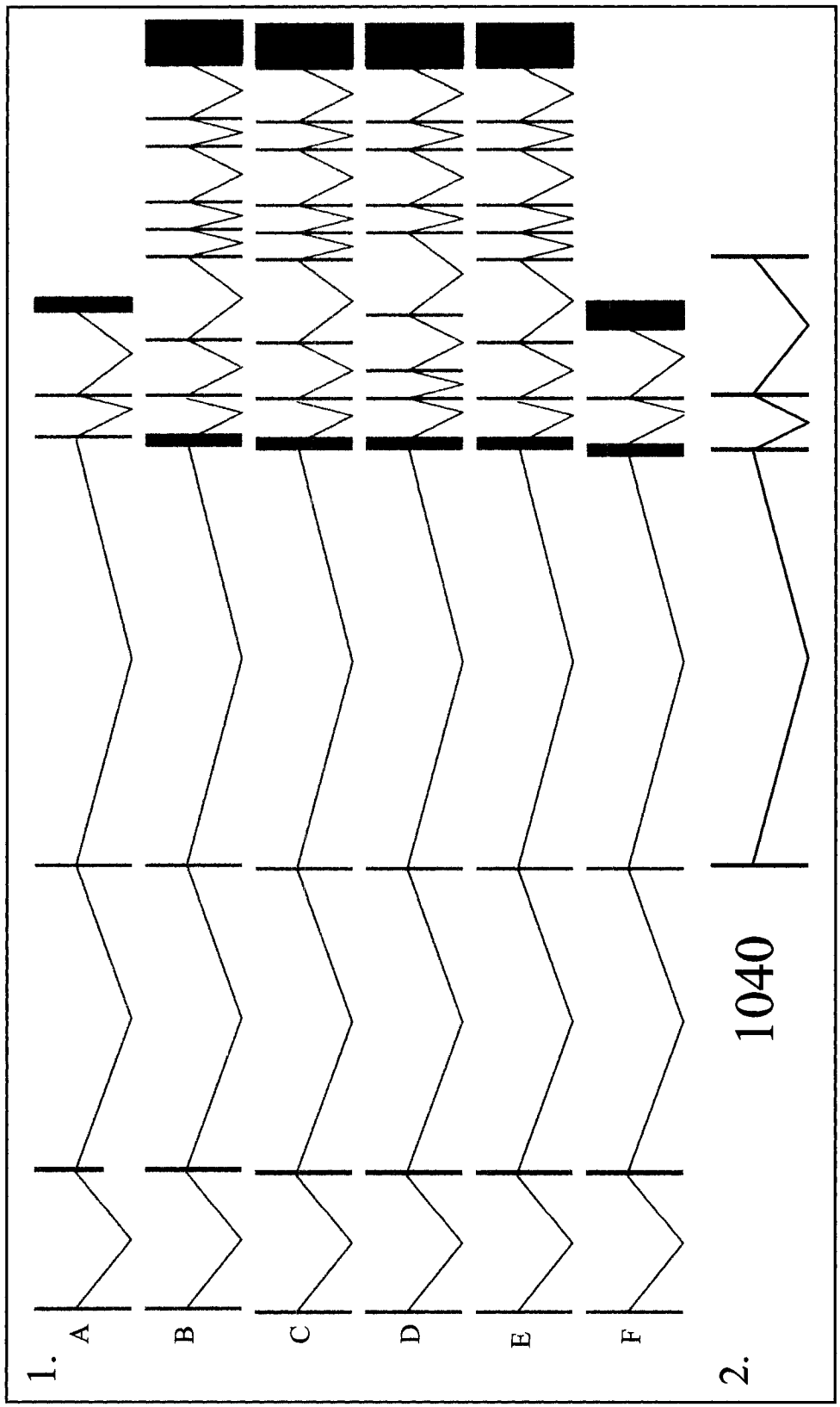
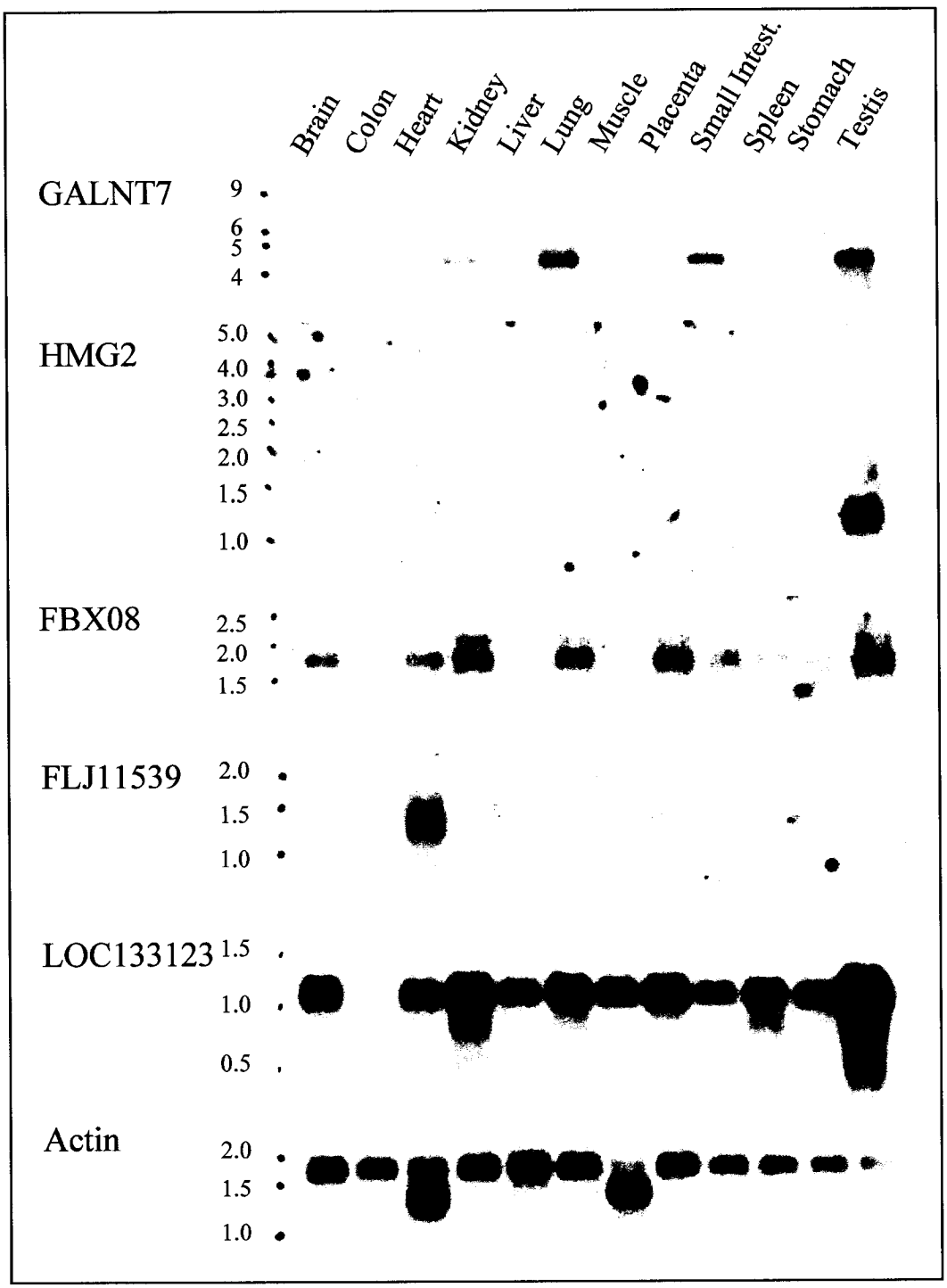

1.
A
B
C
D
E
F
2.
1040

Figure 3.8. Northern Blot analysis of known and predicted genes in the 4q34 region. Northern blots containing several adult human tissues were probed with radioactively labeled *GALNT7*, *HMG2*, *FBX08*, FLJ11539, LOC133123 and Actin respectively. Exposure times varied from overnight to 2 weeks. Blots were stripped and reprobed several times. Actin was used as an RNA loading control. The size markers (kb) are labeled on the left hand side of each blot. ***GALNT7.*** The major *GALNT7* transcript is 4.26 kb, and expressed at varying levels in the included tissues. ***HMG2.*** High expression of a 1.3 kb transcript is seen in testis. Low levels of expression of this transcript are also seen in brain, kidney and lung. ***FBX08.*** The reported major transcript of *FBX08* is 2.0 kb and is seen in most tissues. However, a transcript of approximately 1.7 kb appears to be more abundant in the tissues analyzed here. Most tissues with *FBX08* expression appear to have both 1.7 kb and 2.0 kb fragments. **FLJ11539.** This transcript is predicted to be 1.58 kb. We observed high expression of a transcript of approximately 1.4 kb in heart, with low expression in liver and placenta. A smaller 1.0 kb transcript was observed in testis. **LOC133123.** High levels of expression of a 1.2 kb fragment were seen in all tissue except colon, which had very low levels.


Brain
Colon
Heart
Kidney
Liver
Lung
Muscle
Placenta
Small Intest.
Spleen
Stomach
Testis
GALNT7
9
6
5
4
HMG2
5.0
4.0
3.0
2.5
2.0
1.5
1.0
FBX08
2.5
2.0
1.5
FLJ11539
2.0
1.5
1.0
LOC133123
1.5
1.0
0.5
Actin
2.0
1.5
1.0

3.5.3 High mobility group (nonhistone chromosomal) protein 2 - *HMG2*

HMG2 encodes a member of the high mobility group protein family. These proteins are chromatin-associated proteins distributed ubiquitously in the nucleus and cytoplasm of higher eukaryotic cells [95]. *HMG2* is involved in the bending and formation of DNA circles and in the final step of VDJ recombination. *HMG2* may also be a transcription and differentiation factor [97].

HMG2 covers 3 kb of sequence and was confirmed to be present in the two BACs 798M19 and 297M24 with Southern blotting, as expected from annotation in the available databases (Figure 3.6). *HMG2* encodes 9 predicted splice variants based on mRNA evidence (Table 3.5) with the major transcript being 1.3 kb. Northern hybridization shows high expression of an approximately 1.3 kb transcript in testis and low expression of the same transcript in brain, kidney and lung (Figure 3.8). Low levels of a 1.6 kb transcript are seen in brain, and testis, and a 2.3 kb transcript is seen in heart and brain. All of these transcripts match the sizes of alternative transcripts annotated in the public database. These results could reflect cross-hybridization to *HMG2* homologues, although this was not investigated.

Recently, *HMG2* has been implicated in the pathology of Facioscapulohumeral muscular dystrophy (*FSHD*) through its role in a multiprotein complex. The complex consists of *YY1*, a transcriptional repressor, *HMG2*, an architectural protein and nucleolin. This complex was demonstrated to act as a transcriptional repressor of genes in the 4q35 region. Inappropriate transcriptional de-repression of 4q35 genes is proposed to result in the disease [98].

3.5.4 Sin3-associated polypeptide, 30kD - *SAP30*

SAP30 encodes a protein that is a component of the histone deacetylase complex, which is involved in deacetylating core histone octamers. Histone acetylation plays a key role in the regulation of eukaryotic gene expression. The *SAP30* gene covers approximately 7 kb of sequence and is present in the BACs 789M19 and 297M24 based on *in silico* analysis. Northern blot analysis by Laherty *et al.* indicated that *SAP30* was expressed in many tissues, with low expression in brain [99].

3.5.5 Scrapie-responsive gene 1 - *SCRG1*

Dandoy-Dron *et al.* identified *SCRG1* while looking for increased gene expression in mice infected with the Scrapie protein [100]. *SCRG1* is preferentially expressed in brain. The transcript of the human orthologue of *Scrg1* has been shown to be 0.7 kb covering 78 kb of genomic sequence [101]. Southern blot analysis confirms the localization of *SCRG1* in the BACs 789M19, 297M24 and 489G11 (Figure 3.6). We could not detect *SCRG1* on a multiple tissue Northern blot (data not shown) possibly because *SCRG1* is enhanced in Scrapie-infected brains. Analysis of the DNA sequence of all exons, and intron-exon boundaries of the *SCRG1* gene in our PD family did not reveal any potential disease causing mutations (data not shown).

3.5.6 Heart-and-neuralcrest derivatives-expressed 2 - *HAND2*

HAND2 is a helix-loop-helix (HLH) transcription factor expressed in the vascular mesenchyme and vascular smooth muscle cells. It has been suggested that *HAND2* is

required for vascular development through a Vascular Endothelial Growth Factor (*VEGF*) signalling pathway [102]. The predicted transcript size is 582 bp covering approximately 5.5 kb of genomic sequence. *HAND2* is expressed mainly in the developing heart, and was therefore not identified in the adult heart tissue or any adult tissue represented on our Northern blot (data not shown).

3.5.7 F-box only protein 8 - *FBX08*

F-box proteins are a family of proteins that make up one of the four subunits of ubiquitin protein ligases (SCFs). SCFs bring ubiquitin conjugating enzymes to substrates that are specifically recruited by the F-box proteins [103]. F-box proteins are the largest class of E3 ubiquitin ligase receptors [104]. The family consists of approximately 26 human proteins. *FBX08* is annotated to be present in the BAC 148L24, and this was confirmed with Southern Blotting (Figure 3.6). The *FBX08* transcript is 2.0 kb covering 47 kb of genomic sequence and was observed in all tissues except colon (Figure 3.8). However, a transcript of approximately 1.7 kb appeared to be more abundant than the 2.0 kb predicted major transcript. This could be a result of cross-hybridization to other F-box proteins.

Ubiquitin ligase (*UCH-L1*) has previously been shown to play a role in the pathogenesis of Parkinson's disease in a kindred where two brothers were affected with the disease [105,106,107]. Therefore, *FBX08*, with its implied function in the ubiquitin pathway, was screened as a candidate for Parkinson's disease in our family. Analysis of the DNA sequence of all exons, and intron-exon boundaries of the *FBX08* gene from patient samples did not reveal any potential disease causing mutations (data not shown).

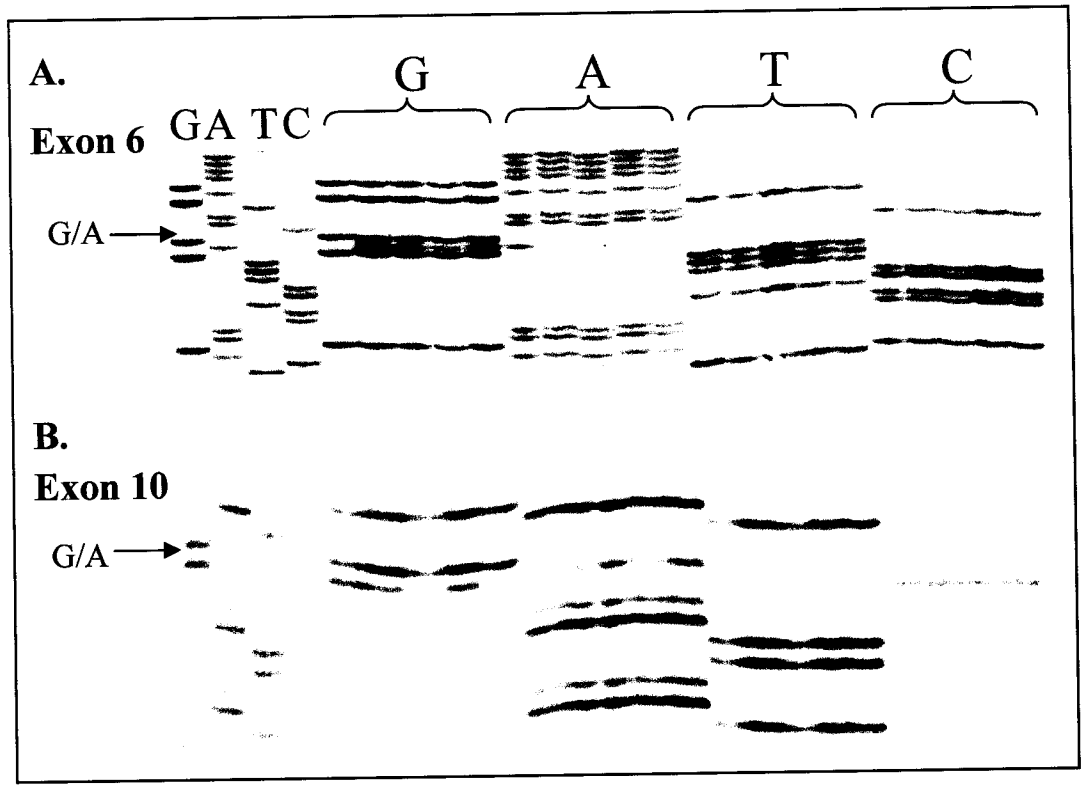
3.5.8 KIAA1712

KIAA1712 is a predicted protein based on a cDNA identified by Nagase *et al.* through a high-throughput screen of a large insert brain cDNA library [108,109]. The predicted transcript size is 4.5 kb and KIAA1712 covers approximately 36 kb of genomic sequence. Expression patterns could not be confirmed by Northern blotting. Functional characterization has not been undertaken, but isolation from brain makes this a strong candidate for Parkinson's disease. Sequencing of the coding sequence of this gene from normal and affected PD family members was completed during the course of this research. Ten coding and 2 non-coding exons were sequenced and no mutations were found in the coding sequence of intron/exon boundaries. Two previously characterized polymorphisms were identified (Figure 3.9).

3.6 Predicted genes in the 4q34 region

At NCBI, a program called GenomeScan has been used to identify potential coding sequences by computer algorithms [110]. This program takes consensus splice sites and other factors into consideration when predicting genes using the published human genome sequence. Other programs, including Aceview, take mRNA and EST evidence into account and predict transcripts from the genomic sequence data [94]. There are numerous predicted genes in the 4q34 region. A list is included in Table 3.6, but the detailed analysis of each gene is outside of the scope of this thesis. Instead, the predicted genes with expression data from Northern blot analysis and those that are represented by cDNA clones shown to map back to 4q34 will be discussed briefly in this section.

Figure 3.9. Polymorphisms identified through sequencing of the 11 exons of predicted gene KIAA1712. The order of patients is: non-affected, affected, non-affected, affected, affected. Patient DNA was amplified in four PCR reactions each containing one of radioactively labeled A, T, C, or G. Reactions were loaded (G,A,T,C of unaffected individual followed by G's of all patients, A's, T's and C's) on a 5 % polyacrylamide gel and electrophoresed for 2-3 hours. Following drying, the gel was used to expose film (KODAK) for 2-3 days. The sequence was read manually and polymorphisms can be seen as variations in patient's sequences. A. A G/A polymorphism is seen in exon 6. B. A G/A polymorphism is seen in exon 10. Both polymorphisms indicated have been previously reported by NCBI [54].


A.
G
A
T
C
Exon 6
GA TC
G/A
B.
Exon 10
G/A

3.6.1 FLJ11539 (formerly LOC152952)

FLJ11539 is a predicted gene based on mRNA and EST clones. The original cDNA clone was isolated from embryonic tissue, mostly head, placenta and neuroblastoma [54]. This predicted gene covers 55 kb of genomic sequence and has 7 putative alternative transcripts (Table 3.5). There is no predicted function for FLJ11539 although some isoforms are thought to be localized to the mitochondria [94].

Oligonucleotide primers were designed from the putative mRNA sequence and Northern blot analysis showed high expression of an approximately 1.4 kb transcript in heart and low expression of the same transcript in liver and placenta. A smaller transcript of approximately 1.0 kb showed a low level of expression in testis. We observed no expression in brain (Figure 3.8). The predicted transcript sizes for FLJ11539 are listed in Table 3.5, however the transcripts seen in this analysis do not overlap with any of the predicted transcript sizes.

3.6.2 LOC133123

LOC133123 is predicted based on similarity to Ribosomal protein L5. Ribosomal protein L5 is highly represented in the human genome. Copies similar to LOC133123 transcript are located on several chromosomes including 1, 4, 5, 8, 11, 15, and 22 (Table 3.7). The fact that the sequence for LOC133123 appears to be spliced into exons in some cases and not in others (represented by two "hits" in Table 3.7) indicates that LOC133123 might often be present as a pseudogene, due to the lack of intronic sequence.

Table 3.7. *In silico* analysis of predicted gene LOC133123.

Location of similarity	Chromosome	E value	Identities
NT_007988	8	e-146	87 %
NT_006257	4	e-117, e-117	100 and 100 %
NT_011520	22	3e-87	92 %
NT_035316	15	4e-86, 6e-85	92 and 93 %
NT_004511	1	3e-83, 1e-67	91 and 88 %
NT_006397	4	2e-81, 2e-52	92 and 85 %
NT_030680	5	7e-81	92 %
NT_035088	11	1e-79	93 %
NT_021979	1	1e-78, 3e-39	94 and 97 %

Note - The Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. Identity refers to the percent of nucleotides that match between the two sequences (LOC133123 and the human genome). Clones with 2 E and 2 identity values have spliced sequence LOC133123.

Strong hybridization to all tissues is seen by Northern blot analysis indicating that this transcript is highly expressed in all tissues tested (Figure 3.8). The high expression seen is likely due to detection of several of the transcripts representing proteins similar to Ribosomal protein L5, not just LOC133123.

3.6.3 LOC256573

LOC256573 is an *in silico*-predicted protein in the 4q34 region with EST support. This transcript is predicted to be approximately 0.7 kb and contains a carbohydrate-binding domain. LOC256573 is located centromeric to *GALNT7* and shows similarity to *GALNT9* (Figure 3.4) [54]. Therefore, this gene is annotated as having been formed from presumed gene triplication [54]. No expression data has been reported for this predicted transcript.

3.6.4 Genes from the Aceview Database

Aceview shows the alignment of mRNAs to the genome sequence, and the genes and transcripts are reconstructed from these alignments using the Acembly program developed by Jean and Danielle Thierry-Mieg [94]. This database also reports similarity to known proteins and conserved motifs, although protein analysis is not used to generate the model transcript. Analysis of Aceview predicted proteins was accomplished entirely using an *in silico* approach. All of the Aceview predicted genes are listed in Table 3.6. Due to the large number of Aceview predicted transcripts, only those that are relevant to the cDNA clones isolated through direct selection are discussed.

Hs4_6414_29_2_3431 is an Aceview predicted transcript that lies within an intron of *GALNT7* and has 3 predicted exons [94]. This predicted gene is defined by one cDNA clone (gi_10437426) and one sequence. There is no expression data for this predicted transcript, although the cDNA was isolated from a colon library [54]. A second Aceview-predicted transcript in the 4q34 region is Hs4_6417_29_3_582. This protein is predicted based on 9 cDNA clones and 10 sequences isolated from tissues including kidney and lung, but not brain [54,94]. There is no known function and the complete transcript has not been identified.

3.7 Clones identified by direct selection

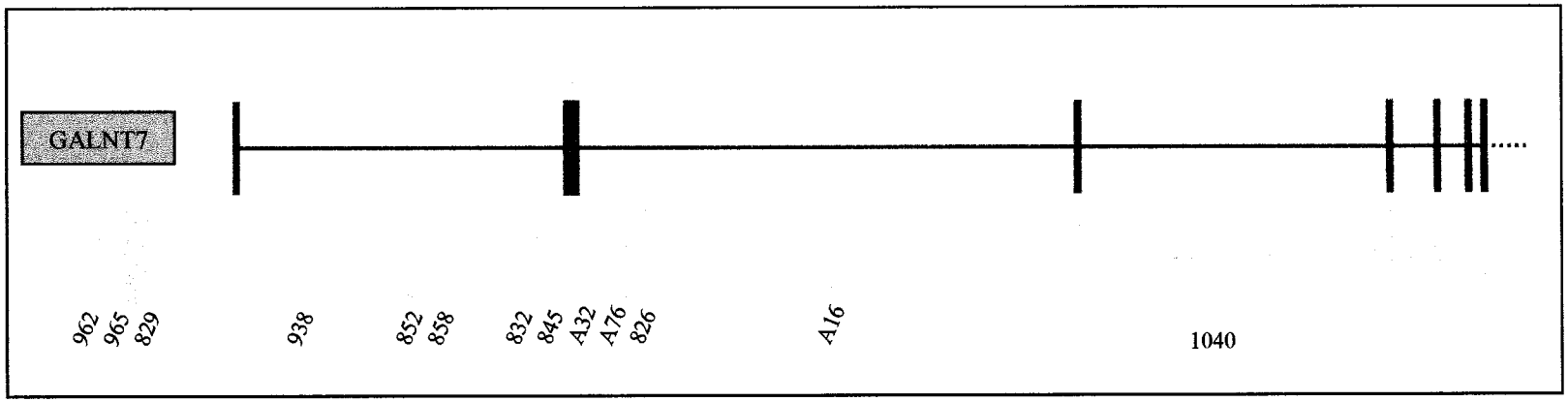
Overall, 351 cDNA clones were sub-cloned and characterized from direct selection at the 4q34 region. Approximately 1/3 of the clones (106) were found to contain inserts greater than 100 bp in length. Sequencing and analysis of clones with inserts was performed. Sequences were compared with those compiled in the public databases. Of these, 55 were found to be greater than 50 % repetitive, 25 mapped directly to the 4q34 region with 17 individual sequences (refer to Appendix 2), 23 mapped to other chromosomal locations and 3 could not be placed on the human genome. Significant similarities or identities of the 4q34 clones determined through *in silico* analysis are listed in Table 3.8. The position of each clone is illustrated on the physical map of the 4q34 region in Figures 3.4 and 3.5. The majority of clones are clustered around the *GALNT7* locus (Figure 3.10). Table 3.9 outlines typical results from clones that did not map to the 4q34 region. When assessing the similarity of clones to genomic sequence, a *P* value of e -15 was considered significant, as described by Roux *et al.* [111]. Additionally, 22 clones

Table 3.8. Summary of the clones mapping to 4q34.

Clone	Size(bp)	Tissue	BAC Overlap	ORFfinder	Aceview	EST
A04	318	testis	10K16, 500A5			
962	187	testis	10K16, 500A5	LOC256573		
965	622	testis	10K16, 500A5			
829	263	testis	10K16, 500A5	AB018478		
938	122	N/A	10K16, 500A5			
852	463	testis	10K16, 500A5			
858	251	N/A	10K16, 500A5			gi_5436877
832	160	N/A	10K16, 500A5			
835	154	testis	10K16, 500A5			
A32	104	placenta	10K16, 500A5		Hs4_6414_29_2_3431	gi_10437426
A76		testis	10K16, 500A5			
826	277	adult brain	10K16, 500A5			
A18	432	testis	10K16, 500A5			
1040	384	testis	500A5, 297M24, 789M19	XM_003527		
944	238	testis	297M24, 789M19		Hs4_6414_29_3_582	gi_4077873
A100	316	N/A	489G11	Ribo protein L5		
833	222	placenta	148L2	KIAA1712		gi_21177337

N/A – Not available. Tissue tags were not observed for some clones, likely due to degradation prior to the sub-cloning step.

Figure 3.10. The clones surrounding the *GALNT7* locus. The centromeric portion of *GALNT7* is shown along with the isolated cDNA clones that map into this region. Twenty-one of the twenty-five clones map into this region of the physical map. Clones that were isolated more than once are represented by only one clone on the map. Clones **832** and **845** are juxtaposed at an EcoR1 site. Clone **1040** represents 4 exons of the *GALNT7 transcript*, although exon 6 is not present. Clone **A32** overlaps with exon 2 of *GALNT7*, however, this does not represent coding sequence in the transcript.



***GALNT7* transcript** with exons (vertica)l and introns (horizontal). The dotted line indicates that only a portion of the gene is illustrated

Complementary DNA clone (vertical box). The dottted line represents intronic sequence that was not present in the isolated clone

Table 3.9. Clones not mapping to the 4q34 region.

Clone(s)	Repeats	Highest Similarity
821, 919, 924, 934, 957, 960, 975, 977, 994, 995, 1007,1030, A29, A30	LINE	
824, 827	SINE	
836	None	None
831	None	Exon in KIAA1463, cromosome 12
840, 911, A23	None	Cytochrome Oxydase
843, A81	Harlequin LTR	
860	Mitochondrial	
941	None	Chromomsome 5
940	None	Chromosome X
948	None	Chromosome 1
951, A55	None	Chromosome 11
966, 968	Simple repeat	PLP1 on chromosome X
A15, A22, A24, A11	Poly A only	
A01	None	Chromosome 9
A43	None	Cromosome 2
A60	None	Chromosome 17
A61	none	Chromosome 6
A96	DNA MER1	

had *EcoR1* sites at one or both ends, detectable because they lacked the RXG oligonucleotide sequence, which was lost during sub-cloning into the *EcoR1* site of pBluescript. This indicates that a significant number of clones represent fragments of cDNAs. At the present time (Jan, 2003) there is evidence linking 7 of the 17 clones to transcriptional units within the 4q34 region as seen by significant similarities to known or putative transcriptional elements in the 4q34 region. A detailed analysis of each of the clones follows.

3.7.1 Clones that map to 4q32 by sequence only

There were 25 clones identified that could be mapped through *in silico* analysis directly into the 4q34 region. However, 18 of these clones (A04, 965, 829, 938, 852, 832, 838, 937, 939, 943, 954, 955, 835, 845, A76, 826, A16, and A18) had no evidence supporting their classification as transcriptional units because *in silico* analysis showed no significant similarity to known or predicted genes, ESTs, or known or predicted proteins, or conserved domains. A list of the tissue from which each clone was isolated, the size and the BAC in which the clone is present is given in Table 3.8. No further data could be identified for these clones.

3.7.2 Clone 962 – similarity to a predicted protein

Clone 962 is 187 bp and originated from testis mRNA. BLAST searches against the nr database localized it to the 4q34 region within the BAC 10KI6 (Figure 3.11). A search of clone 962 for possible open reading frames using ORFfinder revealed that one

Figure 3.11. *In silico* analysis of clone 962. **A.** The sequence overlap of the clone 962 and the BAC 10K16 in the 4q34 region. The 5` and 3` ends do not overlap due to vector sequence in the clone. **B.** The sequence and amino acid overlap between clone 962 and the predicted protein LOC256573 as determined by ORFfinder is shown. **C.** Exon 1 of LOC256573 ends at amino acid 19 (refer to arrow), however, in clone 962, the nucleotide sequence continues without a splice (i.e. contiguous within the genomic sequence). When the entire sequence of clone 962 is translated, it contains several in frame stop codons seen in fuchsia.

A.

```
Clone 962:   19 acatgagggccaatttatgattatgaatttcttagatgatggagtgaatacgtgtagagg 78
BAC 10K16:64811 acatgtgggccaatttatgattatgaatttcttagatgatggagtgaatacgtgtagagg 64870

Clone 962:   79 gagatttctcttcaaaagttgagggaatgagttaattttagttctctctcaccttgggga 138
BAC 10K16:64871 gagatttctcttcaaaagttgagggaatgagttaattttagttctctctcaccttgggga 64930

Clone 962:  139 tacagccctttggaatcctggttttggcagacctatctgcttttcc 185
BAC 10K16:64931 tacagccctttggaatcctggttttggcagacctatctgcttttcc 64977
```

B.

```
214 ttttttttttttttttctttttttttgggaaaagcagataggtctgccaaaaccagga
    F  F  F  F  F  F  L  F  F  W  E  K  Q  I  G  L  P  K  P  G
154 ttccaaaagggctgtatccccaaggtgagagagaactaaaattaactcattccctcaact
    F  Q  K  G  C  I  P  K  V  R  E  N  *  N  *  L  I  P  S  T
 94 tttgaagagaaatctccctctacacgtattcactccatcatctaagaaattcataatcat
    F  E  E  K  S  P  S  T  R  I  H  S  I  I  *  E  I  H  N  H
 34 aaattggccctcatgtgctagatctcgagaattc 1
    K  L  A  L  M  C  *  I  S  R  I
```

C.

```
Clone 962      11 EKQIGLPKPGFQKGCIPK 28
                  EKQIGLPKPGFQKGCIPK
LOC256573      2  EKQIGLPKPGFQKGCIPK 19
```

frame matched perfectly to the predicted protein LOC256573. The overlap between 962 and LOC256573 is illustrated in Figure 3.11. Clone 962 overlaps with exon one of the predicted protein, but does not end at the predicted splice site. The area where the similarity between clone and predicted protein ends occurs at the splice site of the predicted exon 1 (Figure 3.11). There are three in-frame stop codons if the clone is to be translated from start to finish (Figure 3.11). Neither clone 962 nor LOC256573 show any similarity to know proteins.

3.7.3 Clone 858 - EST support

Clone 858 was isolated from unknown tissue (there was no tissue tag) and is 251 bp in length. Clone 858 overlaps with one EST, gi_5436877, which is an image clone of total length 661 bp (Figure 3.12). There is an internal *EcoR1* site at the 5' end of clone 858. The EST does not have similarity to any known or predicted proteins. Translation of 858 by ORFfinder produces reading frames which all include stop codons. However, the +2 frame has stops only in the first ½ of the proposed amino acid sequence (Figure 3.12). Furthermore, a putative start codon is located in the final third of sequence, and this sequence is present in the overlap with the aforementioned EST. It is therefore possible that this clone may represent the 5` portion of an mRNA, including the initiation codon.

3.7.4 Clones A32 and 944 – Aceview predicted gene support

Clone A32 is a very small clone (104 bp) that was isolated from placenta. Clone A32 overlaps with Hs4_6414_29_2_3431, a gene predicted in the Aceview database

Figure 3.12. *In silico* analysis of clone 858. **A.** The overlap between clone 858 and EST gi_5436877. Only the 3` portion of clone 858 overlaps with this EST (overlap starts at base pair 95) The EcoR1 site seen at the 3` end of this clone is underlined in red. **B.** The translation product from the +2 reading frame of clone 858. The portion of clone 858 which overlaps with EST gi_5436877 (as shown in A.) is shown in blue. There are several stop codons (fuchsia colored), however, there are no stop codons in the last 1/3 and a putative start codon (light blue colored) is present. This could represent the 5` portion of an mRNA.

A.

```
Clone 858:       95 gagattcattgatcttcttcactcccagtagtaagttaagcatctaagagacactgtcct 154
EST gi_5436877:   2 gagattcattgatcttcttcactcccagtagtaagttaagcatctaagagacactgtcct 61

Clone 858:      155 ttcaagattccacatgccacccagtgcaccatttactctgtacataacatgggctgtcca 214
EST gi_5436877: 62 ttcaagattccacatgccacccagtgcaccatttactctgtacataacatgggctgtcca 121

Clone 858:       215 tgctgctacacactttcacggag-ggtgaggagaattc 251
EST gi_5436877:  122 tgctgctacacactttcacggagtggtgaagagaattc 159
```

B.

```
   2 aattctcgagatctagtatgg tagtaattaaaaccctaaattatactagcttcttttccc
     N  S  R  D  L  V  W  *  *  L  K  P  *  I  I  L  A  S  F  P
  62 caaaaggaatttttccacttttttttttttttgagattcattgatcttcttcactccca
     Q  K  E  F  F  H  F  F  F  F  F  E  I  H  *  S  S  S  L  P
 122 gtagtaagttaagcatctaagagacactgtcctttcaagattccacatgccacccagtgc
     V  V  S  *  A  S  K  R  H  C  P  F  K  I  P  H  A  T  Q  C
 182 accatttactctgtacataacatgggctgtccatgctgctacacactttcacggagggtg
     T  I  Y  S  V  H  N  M  G  C  P  C  C  Y  T  L  S  R  R  V
 242 aggagaattc 251
```

atg Putative start codon tga Putative stop codon gaattc *EcoR1* site

(Figure 3.13, Table 3.6). This predicted gene is incomplete, and contains a very long 5` UTR [94].

Clone 944 is 218 bp and was isolated from testis. Clone 944 overlaps with the Aceview-predicted gene Hs4_6414_29_3_582 (Figure 3.13, Table 3.6). As already discussed, Aceview predicts genes using EST, and mRNA evidence. Neither of these predicted genes have similarity to any known proteins, and no conserved domains are present. The size of these clones limits our ability to detect expression with Northern blotting.

3.7.5 Clones 1040 and 833 – known transcriptional units

Clone 1040 overlaps with 4 exons of the gene *GALNT7*. The 384 bp of clone 1040 represents exons 3, 4, 5 and 7 correctly spliced, except that the entire exon 6 is absent (Figure 3.10 and Figure 3.14). At the time of this experiment, it was difficult to determine if any splice variants were known for *GALNT7*. Within several months, information was available in the public database indicating that *GALNT7* had 4 alternative transcripts of various sizes. However, clone 1040 did not conform to the splicing of any of these published variants, and it was apparent 1040 represented an alternatively spliced transcript element. In Sept. 2002, the same database which originally indicated the 4 splice variants was updated to include 2 further variants; however, clone 1040 still did not align with any of the proposed transcripts (Figure 3.7).

Clone 833 is 222 bp and isolated from placental tissue. BLAST results show that this clone is 100 % similar to sequence in the BAC 148L24. Clone 833 represents a

Figure 3.13. *In silico* analysis of clones A32 and 944. **A.** The overlap between the nucleotide sequence of clone A32 and the EST gi_10437426 is shown. This EST was used for the prediction of the Aceview mRNA Hs4_6414_29_2_3414. **B.** The overlap between the nucleotide sequence of clone 944 and EST gi_4077873 is shown. This EST was used for the prediction of the Aceview mRNA Hs4_6414_29_3_582. The 5` and 3` ends of both clones do not overlap with the respective EST sequences due to vector sequence in the clones.

A.

```
Clone A32:    14 ctgcagagcgtcactattttgcttaagaaaaagaatctcagtgatgagtctggaagtctg 73
gi_10437426:183 ctgcagagcgtcactattttgcttaagaaaaagaatctcagtgatgagtctggaagtctg
124

Clone A32:    74 aggttactttccagcactt 92
gi_10437426:123 aggttactttccagcactt 105
```

B.

```
Clone 944:       20 gcagggyaagttgagtttccaatcagggcaagaagggtttctgacctataaaaacaatgg 79
EST gi_4077873: 35 gcagggtaagttgagtttccaatcagggcaagaagggtttctgacctataaaaacaatgg 94

Clone 944:       80 ttagctgtaaggcatatataaacttttattaaccttggacctgactcaagtaaaaattat 139
EST gi_4077873: 95 ttagctgtaaggcatatataaacttttattaaccttggacctgactcaagtaaaaattat 154

Clone 944:      140 tacacttatatttactatttaatttatagaagtggcaaactggtgacctttgggcagaat 199
EST gi_4077873:155 tacacttatatttactatttaatttatagaagtggcaaactggtgacctttgggcagaat 214

Clone 944:      200 cttacctatagacatgttttgaccccaaaattggaattc 238
EST gi_4077873:215 cttacctatagacatgttttgaccccaaaattggaattc 253
```

Figure 3.14. *In silico* analysis of clone 1040. The sequence of clone 1040 overlaps exons 3, 4, 5 and 7 of the gene *GALNT7*. Clone 1040 does not include the sequence for exon 7, and therefore may represent a novel alternative splice of this gene. The 5` and 3` ends do not overlap due to vector sequence in the clone.

Exon 3

Clone 1040: 9 gagatctcaaagttttcaccaccccaaatctggagacctggatcatagagacc-aattca 67
GALNT7: 826529 gagatctcaaagttttcaccaccccaaatctggagacctggatcatagagacccaattca 826470
Clone 1040: 68 aagaagaactctcgttcaatggcaaataatccccccagccgtggctggg-acct 119
GALNT7: 826469 aagaagaactctcgttcaatggcaaataatccccccagccatggctggggacct 826417

Exon 4

Clone 1040:116 acctgtccttagatatgggagctacaagtggtgcataccagttaactgccacctcacagt 175
GALNT7: 820217 acctgtccttagatatgggagctacaagtggtgcataccagttaactgccacctcacagt 820158
Clone 1040:176 gggcatcaaggtatatcaaaacct 199
GALNT7: 820157 gggcatcaaggtatatcaaaacct 820134

Exon 5

Clone 1040:196 acctgtccaagtttagccttctgagcaccaatacttcgtgcttgaattaaaccttccctt 255
GALNT7: 819900 acctgtccaagtttagccttctgagcaccaatacttcgtgcttgaattaaaccttccctt 819841
Clone 1040:256 ctttcatttcgaaataccttcactaggccattccacagcttaatatattcatccagtttt 315
GALNT7: 819840 ctttcatttcgaaataccttcactaggccattccacagcttaatatattcatccagtttt 819781
Clone 1040:316 tcttataagtgttct-tta 333
GALNT7: 819780 tcttttaagtgttctgtta 819762

Exon 7

Clone 1040:329 ctttattactgaaatcgtcaattaacacaatttctgcta 367
GALNT7: 816646 ctttattactgaaatcgtcaattaacacaatttctgcta 816608

portion of exons 2 and 3 from the predicted protein KIAA1712 (Figure 3.15). There are EST similarities to those ESTs which are annotated for the KIAA1712 protein, for example gi_23711492, gi_22679553 and gi_21177337.

3.7.6 Clone A100 – a pseudogene?

Clone A100 is 316 bp and there was no tissue tag so tissue origin could not be determined. BLAST results show that this clone is 92% similar to sequence from BAC 489G11 (Figure 3.16). This indicates that the clone is actually from somewhere else in the human genome, but was isolated due to the similarity to the 4q34 sequence. The highest similarity of clone A100 is 95% similarity to sequence from chromosome 22. 94 % similarity is shown to chromosome 1 and 93 % similarity to chromosomes 15, and 1 (Table 3.10). Clone A100 overlaps with the predicted protein LOC133123, which is highly represented in the human genome (Table 3.10).

Figure 3.15. *In Silico* analysis of clone 833. Clone 833 overlaps with exon 3 of the predicted transcript KIAA1712. The 5` and 3` ends do not overlap due to vector sequence in the clone.

```
Clone 833:275 tattaggcagatgggattggctggtctgtgctctgctaggtgaaaaattagacgatttca 334
KIAA1712: 191 tattaggcagatgggattggctggtctgtgctctgctaggtgaaaaatttgacgatttca 132

Clone 833:335 agaattggagctgaatctgatgttaagtaatggcaacaggtgacttaaaaagaagcttac 394
 KIAA1712:131 agaattggagctgaatctgatgttaagtaatggcaacaggtgacttaaaaagaagcttac 72

Clone 833:395 ggaacctagaacaggtgctccgcttgctaaattatcctgaagaggtggactgtgt 449
KIAA1712:  71 ggaacctagagcaggt-ctccgcttgctaaattatcctgaagaggtggactgtgt 18
```

Figure 3.16. *In Silico* analysis of clone A100. Clone A100 overlaps with the sequence from BAC 489G11, however, only with 92 % similarity. The base pair differences between the two sequences can be seen in red. The 5` and 3` ends do not overlap due to vector sequence in the clone.

```
Clone A100:    34 aaagaacagcgtaactccagacatgatggaggagatgtataagaaagctcatgctgctat 93
BAC 489G11:27980 aaagaacagcgtaactccagacttgatggaggagatgtataagaaagctcatgctgctgt 27921

Clone A100:    94 acgagagaatccagtctatgaaaagaagcccaagaaagaagttaaaaagaagaggtggaa 153
BAC 489G11:27920 acgagagaatccagtctatgaaaagaagcccaagaaagaaattaaaaagaagaggtagaa 27861

Clone A100:   154 ccgtcccaaaatgtcccttgctcagaagaaggatcgggtagctcaaaagaaggcaagctt 213
BAC 489G11:27860 ctgtcccaaaatatcccttgctcagaagaaagat tgggtagctcaagagaaggcaagctc 27801

Clone A100:   214 cctcagagctcaggagcgggctgctgagagctaaacccagcaattttctatga-- ttctt 271
BAC 489G11:27800 cctcagagctcaggagcgggctgccgagagctaaaccaaacaattttctatgaggatttt 27741

Clone A100:   272 tcagatatagataataaacttatgaac 298
BAC 489G11:27740 tcagataaagacaataaacttatgaac 27714
```

Table 3.10. BLAST similarity of clone A100 to the human genome.

Location of similarity	Chromosome	E value	Identities
NT_011520	22	e –115	95 %
NT_028050	1	e –115	94 %
NT_010178	15	e –108	93 %
NT_004686	1	e –104	93 %
NT_004487	1	e –103	92 %
NT_006257	4	e –102	92 %
NT_006520	5	4e –98	92 %
NT_010332	15	6e –97	91 %
NT_009151	11	6e –94	91 %

Note - The Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. Identity refers to the percent of nucleotides that match between the two sequences (A100 and the human genome).

Chapter 4

4.0 Discussion

4.1 Genes in the 4q34 region – additional information

To assess the possibility that some of the known and predicted transcripts within the 4q34 region might be possible candidate genes for PD, a brief analysis of the known and predicted genes in the 4q34 region was undertaken. Southern blot analysis confirmed the location of the known genes *GALNT7, HMG2, FBX08, KIAA1712* and *SCRG1* and the predicted genes LOC152952, LOC133123, LOC166847, LOC152950 and LOC166475 within the BACs of the 4q34 region (Figure 3.6).

Three known genes, *GALNT7, HMG2,* and *FBX08* were analysed by Northern blotting to assess their expression patterns. Analysis of *GALNT7* showed expression in brain, liver, lung, small intestine, and testis of a transcript of approximately 4.3 kb (Figure 3.8). These tissues had not been analysed in the previous characterization [96]. The expression seen in brain, although comparatively low with respect to other tissues on the Northern blot, supports *GALNT7* as a potential candidate for PD even though the proposed function of *GALNT7* does not appear to be related to disease manifestation.

Analysis of *HMG2* identified a transcript of approximately 1.3 kb with high expression in testis and low expression in lung, kidney, and brain. A second transcript of approximately 2 kb was seen in testis and brain (Figure 3.7). A transcript of about 2.4 kb was seen in both brain and heart (Figure 3.8). It has been predicted that *HMG2* has 9

possible alternative transcripts based on known ESTs, but this was not validated by direct experimentation [94]. Several of these transcripts appear to be expressed at low levels as seen on our Northern blot (Figure 3.8).

Northern blot analysis of *FBX08* identified expression of a transcript of approximately 1.75 kb in brain, heart, kidney, lung, placenta, small intestine, spleen, stomach and testis. A transcript of approximately 2.0 kb was also observed in brain, kidney, lung and testis. The annotated *FBX08* transcript is 2.0 kb, however, the expression of the smaller 1.75 kb transcript was more ubiquitous in our analysis. This may reflect cross hybridization to other F-box only proteins.

Northern blot analysis identified expression of two predicted proteins in the 4q34 region. The predicted protein LOC133123 is based on similarity to Ribosomal protein L5 which is highly represented in the human genome (Table 3.7). Southern blot analysis identified that this gene was present in BACs 789M19 and 489G11, this was confirmed with *in silico* analysis.

Southern blot analysis of the predicted gene FLJ11539 confirmed the sequence placement in BAC 489G11 as predicted by the current physical map at NCBI. Northern blot analysis showed high expression of a transcript of approximately 1.4 kb in heart, and low expression in liver, and placenta. Expression of a 1.0 kb transcript was seen in testis (Figure 3.8). The smallest predicted transcript for FLJ11539 is 1.58 kb (Table 3.4), which is larger than those seen in this analysis. The lack of expression in brain lowers the priority of this predicted gene as a potential candidate for PD, however, the expression shown mainly in heart was a novel finding.

4.2 Direct selection in the 4q34 region

4.2.1. 4q34 clones – global analysis

Through the course of direct selection at the 4q34 region, 351 clones were subcloned and analysed. Approximately 1/3 of the clones (106) contained inserts greater than 100 bp. Of these clones, roughly 1/2 (55) were found to be greater than 50 % repetitive, 1/4 (25) could be localized to the 4q34 region and the remaining clones could be localized to chromosomal regions outside of the 4q34 region (23/26) or could not be mapped by sequence to the human genome (3/26). These results are similar to those obtained in a comparative direct selection by Roux *et al.* who identified 17 positive clones from a screen of approximately 300 in which they also found about 50 % of clones were repetitive [111]. However, the results can vary quite widely. For example, in a screen of only 47 clones obtained by direct selection, Tambini *et al.* found that at least 33 (>70%) mapped back to their originating chromosome 7 BAC [76]. Rommens *et al.* report identifying between 4-30 % repetitive clones in various direct selections they have undertaken [80]. The size of the candidate region, the nature of the region (% repeats for example) and the type of vector used as the source of genomic DNA (YAC or BAC) will affect the results of the cDNA selection. In addition, low-copy repeat elements that are transcribed, pseudogenes or genes that are members of gene families will contribute to the population of clones that do not map appropriately to the region of interest [80].

The availability of databases such as nr, HTGS, dbEST, and Aceview and database mining tools like BLAST, BLASPp and ORFfinder, allowed quick and powerful *in silico* analysis of the cDNA selected clones to identify similarities between clone

sequences and known mRNAs, cDNAs and ESTs. Complementary DNA clones were isolated that could be mapped to all BACs, except for the BAC 248N22 which contains no genes and few ESTs. BACs 489G11 and 148L24 had only one clone each identified within their sequence. BACs 10K16, 500A5, 297M24 and 789M19 had several clones represented in each of their sequences (Table 3.8). The majority of clones (23/26) mapped to a region surrounding the known and characterized gene *GALNT7* in overlapping BACs 10K16 and 500A5 (Figure 3.10). Eleven individual clones are clustered within the *GALNT7* transcription unit. However only one, clone 1040, overlaps with transcribed sequence from *GALNT7* (Figure 3.10). The identification of clone 1040 provided us with a positive control in that it showed that transcribed sequences could be isolated by direct selection from the 4q34 region.

GALNT7 has been reported to be ubiquitously expressed at very high levels [94]. The high transcriptional activity in this region may partially account for the high number of cDNA clones that mapped to this area of 4q34. Furthermore, the existence of six alternative transcripts for *GALNT7* may indicate that some of the clones isolated within the *GALNT7* transcriptional unit may represent further transcriptional elements for this gene. EST and mRNA evidence is available for clones 829, 858, and A32 which all map into the *GALNT7* region (Figure 3.10). *In silico* analysis of these clones and the others in the *GALNT7* region, did not show that they overlapped with any of the alternative transcripts of *GALNT7*. However, these clones could represent tissue specific transcriptional variants of *GALNT7*. Further experiments would need to be designed to address this hypothesis. For example, PCR utilizing oligonucleotide primer pairs designed within both the aforementioned clones and *GALNT7* exons could identify

transcripts that are amplifiable from our original cDNA pools. If transcripts could be amplified utilizing an oligonucleotide designed from clone sequence and one designed from *GALNT7* sequence, it would provide evidence supporting the existence of further tissue specific *GALNT7* transcripts. Alternatively, the cDNA clones could be used as probes to isolate larger cDNA fragments from the original cDNA pools. This may be beneficial to determine if these clones do actually represent transcriptional units, especially if multiple exons were identified from individual clones. Furthermore, the fragments would be large enough to successfully screen Northern blots to discern tissue expression patterns. However, more evidence linking the clones to the pathogenesis of PD would be essential before extensive characterization was undertaken.

The analysis of the tissue tags identifies that only one clone was isolated from brain and the majority of clones were isolated from testis and placental cDNA (Table 3.8). The fact that most clones were not identified from our tissue of interest (brain) lowers the priority for further characterization. The lack of tissue tag, seen in several clones was likely due to degradation of the cDNA prior to the sub-cloning step.

Strong correlations have been observed between repeat content, G/C content and gene density [11,12,30,32]. The highest gene densities are observed in H2 and H3 isochores, which are defined as having G/C contents of 43-48 % and >48 % respectively [32]. The G/C content of the BACs representing the 4q34 region is in the range of 38-42 % (Table 3.4, excluding the BACs not used in the cDNA selection). Therefore, this 4q34 region falls into the category of the gene-poor L isochore, which is defined as containing <43 % G/C [32]. This is not surprising because chromosome 4 has been identified as one of the chromosomes with the lowest gene density and also one containing the fewest H3

isochores (gene-rich bands) [12]. The BACs used for cDNA selection appear to have similar G/C contents, in the range of 40 % although the telomeric BACs 248N22 and 148L24 are slightly lower, near 35 % and 489G11 is the highest at ~42 % (Table 3.4). The percent of repeats also appears to be lower in the 3 BACs 489G11, 248N22 and 148L24, from which very few clones were identified (Table 3.4). Furthermore, the percentage of Alu repeats alone, which has been correlated to gene density [12], appears to be lower in these BACs, even though the difference is negligible in 489G11 (Table 3.4). Therefore, the existence of very few known genes or ESTs and the relatively low percentage of Alu repeats in the BACs 248N22 and 148L24 indicates that in fact the most telomeric portion of this region is gene poor. The fact that only one cDNA clone (833) was isolated from these two BACs, and that it represents sequence from a previously identified gene (KIAA1712) supports the notion that this region likely has few transcriptional elements within it. Therefore, identification of novel transcriptional units from this region would be unlikely.

The BAC 489G11 has several predicted genes and 2 known genes represented within it, so the low recovery of cDNA clones from this region cannot be explained by BAC composition alone. The 3 genes in BAC 489G11 which have strong supporting evidence are *SCRG1*, FLJ11539 and *HAND2* (Figure 3.4). The fact that no cDNA clones were isolated that represented sequence from any of these transcripts is most likely explained by the restricted expression patterns of these genes. *SCRG1* expression appears to be up regulated in Scrapie-infected brains [100]. Presumably, the cDNA pool used in this selection was prepared from an uninfected individual, making the identification of *SCRG1* protein less likely. Furthermore, *HAND2* is reported in the literature to be

expressed mainly in the developing heart [102] and we showed that FLJ11539 expression was also largely limited to heart (Figure 3.8). Therefore, the scope of tissues utilized to construct the cDNA pools for direct selection likely did not include those that would facilitate the isolation of transcriptional units from the genes represented in the BAC 489G11.

Results from extensive *in silico* analysis identified two clones containing multiple exons from known genes in the 4q34 region, *GALNT7* and KIAA1712. Additionally, 5 clones can be supported as transcriptional units based on similarities to known ESTs, cDNA clones and predicted mRNAs. The remaining 19 clones could not be shown to match any known gene, protein or conserved sequence in the current databases. The post-transcriptional splicing of mRNAs to remove intronic sequence makes it possible that some intronic sequence may be present in the cDNA pools. However, it is also possible that these clones represent tissue or developmentally specific transcripts that will be difficult to identify. With continued analysis of the public databases, evidence may become available that will support more of the clones as transcriptional units from the 4q34 region.

4.2.2 Clones with similarity to known genes

Seven clones were isolated that could be mapped to the 4q34 region, and had evidence supporting their classification as transcriptional units such as similarity to EST or cDNA sequences. The isolation of clones that represent known transcriptional units was essential for the validation of the direct selection technique within the context of our attempt at the 4q34 region. Furthermore, it provides the opportunity to identify additional

information about these transcripts from a non-biased point of view. In other words, we could potentially identify new transcriptional variants or novel expression patterns based on the selected cDNA utilized in the selection protocol.

In silico analysis of clone 1040 showed that the 384 bp of sequence in this clone aligned with 4 exons of *GALNT7*. Interestingly, the exons were spliced together at the documented splice sites, however, the exons represented in clone 1040 included 3, 4, 5 and 7, exon 6 was missing entirely. Six alternative transcripts for *GALNT7* have been identified in the databases, but none overlap with the transcript represented in clone 1040 (Figure 3.7). Therefore, it is proposed that the isolation of clone 1040 identified a known gene, but also a further alternative transcript of this gene which has not been annotated. The amplification of this transcript from the cDNA library or mRNA representing the same tissues as the cDNA library would ultimately identify if clone 1040 actually represents a novel alternative transcript.

Clone 833 was shown to have similarity to the known protein KIAA1712. This gene was originally identified by Nagase *et al.* in their high-throughput screen of a large insert brain library [90]. The fact that KIAA1712 is a relatively uncharacterized 4.5 kb transcript that is expressed in brain makes KIAA1712 interesting with respect to a neurodegenerative disorder, such as PD. Due to the attractiveness of KIAA1712 as a candidate for PD, DNA sequencing of all exons and intron/exon boundaries was undertaken utilizing PD patient DNA in an attempt to identify any mutations in the coding sequence. The analysis of 9 coding exons and 2 non-coding exons (including surrounding intron-exon boundaries) of 3 PD patients and 2 unaffected controls identified

2 previously characterized polymorphisms. No polymorphisms were found to be segregating with the disease.

4.2.3 Clones with EST, mRNA or cDNA supporting evidence

Clone 962 showed similarity to the predicted gene LOC256573. This predicted gene LOC256573 is a small, 0.7 kb transcript that is predicted based on similarity to *GALNT9* which was postulated to arise from a gene triplication. Clone 962 shows similarity to the first exon of this predicted gene, although the sequence in clone 962 extends past the predicted end of exon 1 (Figure 3.11). Furthermore, the translation of these additional bases creates several stop codons that would be in frame if this is the correct reading frame of the transcript (Figure 3.11). This might indicate that this is actually a portion of a 5' untranslated region. It could also represent a cDNA that was generated prior to the correct post-transcriptional splicing which therefore includes some intronic sequence, or it could represent a pseudogene that will never be translated into a protein. Nonetheless, the isolation of this transcript illustrates that the direct selection technique can potentially be utilized to confirm the identity of *ab initio* predicted transcripts.

Clone 858 was shown to have similarity to a known EST, gi_5436877 isolated from prostate. Translation of clone 858 into protein sequence utilizing the program ORFfinder identifies one reading frame with a putative start codon (Figure 3.12). There are missense/stop codons prior to this start codon indicating that this can only represent the 5' exon. The generation of the cDNA in this direct selection protocol is biased towards identifying 3' sequences because polyadenylated primers are used for first strand

synthesis of cDNA. However, clone 858 contains an *EcoR1* site indicating that the 3' portion of this clone was lost during sub-cloning into pBluescript.

Clone A32, although only 78 bp was shown to align to the predicted protein Hs4_6414_29_2_3431. This protein was predicted in the Aceview database and is based solely on the cDNA clone gi_10437426, isolated from colon [94]. Therefore, clone A32 provides additional support to this predicted protein, which now has two supporting cDNAs. Clone 944 also shows similarity to a predicted protein in the Aceview database, Hs4_6414_29_3_852. This predicted gene is supported by 9 cDNA clones, mainly isolated from kidney. The relatively small amount of evidence supporting these predicted proteins may be a reflection of the expression pattern of the proteins. For example, these proteins may be expressed at a developmentally specific time or in a very tissue specific fashion, which would limit the evidence identified in high throughput screens of cDNAs and ESTs. Therefore, the additional evidence provided by the isolation of clones A32 and 944 provides additional support for the existence of transcripts that have been characterized by *in silico* analysis of the mRNA evidence [94].

Clone A100 shows similarity to the predicted protein LOC133123. As already discussed, this protein is predicted based on similarity to Ribosomal protein L5 which is highly represented in the human genome (Table 3.7). The most significant similarity in analysis of clone A100 is to chromosome 22 with 95 % (Table 3.10). Similarity to chromosome 4 is 92 %. This indicates that in fact this clone likely does not represent sequence from 4q34, but was isolated because of the high similarity to this common ribosomal protein located within this region.

4.3 Direct Selection – Limitations

The cDNA selection technique has been shown to be a fast and powerful method for the isolation of transcribed sequences from a given region [75,76,80,111]. However, several difficulties were encountered in the initial experiments. The most notable problem encountered was the isolation of repetitive sequences. The fact that more than 50 % of the human genome is composed of repetitive elements makes it extremely difficult to undertake direct selection without detecting repetitive sequences. Limiting the isolation of repeats via blocking during hybridization and then screening for repetitive elements following sub-cloning can limit the amount of time spent characterizing repeat-containing clones, but it still means that the characterization of a large number of clones is necessary to identify transcriptional units above the inherent background level of repetitive clones. Furthermore, some transcriptionally pertinent clones will inherently contain repetitive elements, and may therefore be difficult to identify if clones containing repetitive elements are excluded from the analysis.

Another difficulty encountered with this protocol is one that stems from the dynamic nature of the published human genome databases. Currently the published human genome sequence is in draft form. Although estimates are given that predict the finished draft will be available in the spring of 2003, the definition of "finished" in the case of the human genome, does not include regions that are difficult to sequence. Therefore, some important chromosomal regions, and transcriptional units are not currently annotated in the "finished" portion of the human genome sequence. This is illustrated by the fact that less than 90% of known genes can be placed on the current maps, even though estimates of the genome's completeness are higher than 99%. In our

screen of 4q34, three cDNA clones were identified that could not be placed on the map. These sequences could represent sequence that have not yet been placed on the human genome, and may therefore be unidentified transcriptional elements. The fact that so few clones could not be placed, indicates that the 4q34 region is relatively complete in the status of its sequence. This is reflected in the fact that most BACs in the region are in finished status. However, not all regions of the genome are identified as finished. For example, the 4q32 region still contains several BACs with Draft status. It will be essential to re-evaluate these clones that cannot be placed on the map, in order to assess if they are present on BACs that are newly placed in the human genome. It is possible that they illustrate that a portion of the 4q34 map is missing. These clones could also represent contaminating DNA from other sources, however, sequence from bacteria or yeast would be identified when the clone is searched against the nr database.

The nature of the cDNA pools utilized for the direct selection provides a limitation to this procedure. Direct selection with a narrow range of tissues, for example in our case brain, placenta and testis derived cDNAs, will limit the isolation of transcripts to those expressed in these tissues. In our case, this limitation was meant to work to our advantage, because we wanted to identify transcripts expressed in brain, as a means to enrich for the identification of candidate genes for PD. The addition of the testis and placental cDNAs widens the scope of the analysis to include transcripts that may not be expressed in brain. However, this also provides the benefit of ensuring that some results are obtained in the case that there may be no transcripts in the region expressed in the tissue of interest. In that case, it might be deemed that the selection itself was unsuccessful because no transcripts were identified but in fact the negative result was

more likely due to either the narrow scope of tissue used in the analysis, or the paucity of genes in the region.

Finally, the characterization of the large number of clones identified by direct selection may represent a problem. In our case, 17 individual putative transcripts were identified from 351 subcloned sequences. There is the potential to waste extensive amounts of time analysing clones that are not relevant to the region of interest (repetitive, rRNA etc.). Clone characterization must be prioritized based on the context of the direct selection. Clones containing greater than 50 % repeat where not further characterized. Furthermore, we are looking for putative PD disease-causing genes, and therefore we enriched for clones that represent transcriptional units expressed in brain. Only one clone was identified from brain. This clone therefore will have high priority for further characterization. However, the initial characterization of these clones to simply locate them at the 4q34 region required significant time. Analysis of the expression and identities of each clone with Northern blot analysis or RT-PCR will require a large amount of work. Due to the relatively small size of the 4q34 clones, it would be advantageous to initiate experiments to identify full-length cDNAs from size-selected libraries because larger clones provide greater power in database searches. A large insert brain library is available in our lab. Although the human genome is in the final stages of sequencing, significant amounts of data are still being added to the public databases on a daily basis. Therefore, the continued analysis of each clone, especially those that have little or no evidence outside of sequence similarity to 4q34, might eventually identify additional evidence supporting these clones as transcriptional units.

4.4 Initial direct selection at the 4q32 region – a negative result?

Direct selection was used in an attempt to identify novel transcriptional units from the 4q32 region, which lies within the PD locus. Approximately 650 clones were sub-cloned and analysed. In general, the results were not positive because the majority of clones were found to contain repetitive sequences (LINEs, SINES, LTR) or genes that were found on several chromosomes, and were therefore deemed to be low-copy repetitive elements. However, 4 individual clones were isolated that could be mapped back to the 4q32 region. Clones 5 and 41 were mapped to the 4q32 region by Southern blotting while clones 1, 41 and 383 were mapped to the region by *in silico* analysis.

Clone 5 maps to the BAC 82A1 with Southern blotting (Figure 3.3). However, *in silico* analysis places clone 5 on chromosomes 2. Analysis of this clone with Repeatmasker to identify repetitive elements shows that clone 5 is 30 % type MER1 repeat. Clone 41 maps to the BACs 82A1 and 694K14 with Southern blotting (Figure 3.6). *In silico* analysis also maps this clone to the BAC 694K14 (Figure 3.3). The BACs 82A1 and 694K14 do not overlap in the current physical maps. When analysed with the Repeatmasker webserver, clone 41 is shown to contain 21 % MIR3 repeat.

The clones 1, 41 and 383 all map within a small region on the BAC 694K14, within the final exon of a predicted gene DKFZp566D234 when analyzed *in silico* (Figure 3.3). Southern blotting was not undertaken for these clones. Clone 1 contains sequence that is entirely represented in the clone 383. Clone 41 does not overlap with clones 383 or 1. The Repeatmasker webserver identified that clone 383 is 49 % LINE

sequence, clone 41 is 21 % MIR3 repeat, which is a type of SINE repeat, and clone 1 is 58 % LINE repeat. The overlapping sequence in clones 383 and 1 includes the repetitive element and some unique sequence.

The realization that all of these clones contain some portion of repetitive DNA sequence has several implications. Firstly, the repetitive element in the clone likely causes each clone to have a higher probability of being isolated in the direct selection protocol. Repetitive sequences are by definition found in high copy number within the human genome, and therefore within the BACs utilized for direct selection. As a consequence, those clones containing repetitive elements are more likely to be isolated. The result from these 4 clones illustrates this point very well. Clone 5 maps to the 4q32 region by southern blotting, but to an entirely different chromosome by *in silico* analysis. Therefore, it is very likely that the repetitive elements in this clone caused it to be isolated solely by similarity to repetitive elements in the BACs utilized for direct selection, and the unique sequence from the other chromosome was isolated simply by association.

However, it can be seen in the cases of clones 41, 1 and 383 that the repetitive elements gave these clones a selective advantage, but that in fact part of the sequence is unique to the 4q32 region. It is likely that these clones hybridized to the exact region matching 100 % of the sequence (both repetitive and unique elements) rather than hybridized to any repeat that was similar to the repeat in the clone. This is supported by the fact that three individual clones were isolated in the same region of 4q32, making it very unlikely that all of these clones were identified by random chance alone. The possibility still remains that these clones may represent transcriptional elements within

the 4q32 region. However, little supporting evidence was found utilizing *in silico* analysis, thereby indicating that the repetitive elements may have provided a enough selective advantage to make isolation of "junk" DNA possible.

Initially, an experiment was designed as a positive control that would identify a known gene on chromosome 4 that was expressed in brain to verify and optimize the protocol in our laboratory. However, the positive control gene that was chosen was originally mapped through *in silico* analysis to the positive control BAC, but was subsequently found to be present at another chromosomal location. Therefore, we did not identify any transcriptional units from this gene in this particular direct selection experiment because in fact the gene was not located in the BAC utilized. However, this initial attempt illustrated the type of results that are seen in a direct selection where there are no transcriptional units to be found. In further direct selections at the 4q32 locus, very few clones could be mapped to the originating BACs. The same type of results were being observed as those obtained from the original selection experiment. As the selection protocol is PCR-based, there is and innate level of background which will be attained.

In the case of 4q32, almost all of the cDNA clones appeared to be background, as they did not map to the originating BAC clones. All clones that were analyzed contained some proportion of sequence that could be identified as repetitive sequences or low-copy repeat sequences. Even the clones that could be mapped into the 4q32 region were shown to contain at least 20 % repetitive element. However, it is clear that there was some success with the protocol itself, because some clones could be mapped to the region. It can be speculated that the unsuccessful results of these screens were not due to the unsuccessful attempt at the protocol itself, but the nature of the BACs used in the screen.

It has previously been established that the nature of the BAC sequences has a bearing on the results that will be obtained in a direct selection attempt [80]. Analysis of the current sequence of this region in the human genome databases indicates that this region is gene poor. I propose that these results support the notion that there are very few genes in the 4q32 region. The clones that were retrieved represent a level of background that is seen when there are no genes in the region to identify transcripts from. Similar repetitive elements were seen in the 4q34 screen, and may be due to the low gene density in the telomeric BACs in that region. Furthermore, even if there were genes in a region from which to identify transcripts, if the tissues used in the selection protocol are not representative of the expression pattern of the genes in the region, similar results to those presented here would likely be seen.

Therefore, the results for the 4q32 region are not entirely negative. Firstly, the results illustrate that the region was in fact truly gene poor, as was already predicted in the databases. Additionally, through this screen, we were able to optimize our protocol for the direct selection procedure and apply it to a more gene dense region and obtain positive results in a relatively short period of time. The knowledge of the nature of the DNA in the 4q32 region allowed us to choose a region that was more likely to be fruitful in the identification of transcriptional units, namely due to the presence of ESTs and cDNA clones that were uncharacterized.

4.5 Direct selection for identification of novel genes in a candidate region

Positional cloning and computational gene prediction techniques facilitate the identification of genes for disease causing mutations in a candidate region. Our lab has

established linkage of autosomal dominant Parkinson's disease in a large French-Canadian family to chromosome 4q32-q34. This region has not been previously reported to be linked to Parkinson's disease, and therefore represents a novel Parkinson's locus. Identification of candidate genes within the linkage region is essential for identifying the disease-causing mutation. Two known genes within the linkage region, *FBX08,* and *SCRG1* had been excluded as candidates by mutational analysis with DNA sequencing prior to the initiation of this research. Mutational analysis of the gene KIAA1712 identified no mutations segregating with the disease, and it can therefore be excluded as a candidate as well.

To identify further candidate genes for the PD causing mutation, identification of novel genes was undertaken. Simultaneous utilization of both direct selection and *in silico* analysis was shown to be an effective tool for identifying potential novel transcriptional units and rapidly placing them on the human genome map. Twenty-six clones from the 4q34 direct selection attempt were mapped back to the 4q34 region utilizing the available public databases and the database mining tools. The availability of the human genome sequence on the World Wide Web provides a powerful tool for identification of unknown sequences and placement of them on the current genome physical map.

The characterization of all of the clones identified in this screen is well underway. Several clones had evidence prioritizing their characterization with respect to the identification of genes for Parkinson's disease. Methods to identify full-length cDNAs that overlap with the selected clones will help to elucidate if the isolated clones truly represent transcription units. Expression analysis with a variety of tissues including those

that are relevant to PD with techniques like Northern blotting or RT-PCR will help to identify which clones truly represent transcriptional units.

4.6 Conclusions

There is not one methodology for transcript identification that can consistently uncover all of the transcriptional units in any given genomic region. The results of this study illustrate the importance of concurrent analysis of the region of interest with several gene identification methods, including both computational and experimental techniques. The discovery of novel transcriptional elements will be greatly aided by the availability of an accurate and complete human genome sequence. Presumably the majority of transcription units have already been identified and placed on the human genome map, however tissue specific or developmentally regulated expression of some genes causes them to elude current computational programs, and be under-represented in cDNA and EST databases.

References

1 Boguski, M.S. (1999) Biosequence exegenesis. *Science*, **286**: 453-455.

2 Sanger, F. and Coulson, A.R. (1973) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**: 441-448.

3 Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.*, **162**: 729-773.

4 Strauss, E.C., Kobori, J.A., Siu, G., and Hood, L.E. (1986) Specific-primer-directed DNA sequencing. *Anal. Biochem.*, **154**: 353-360.

5 Gocayne, J. *et al.* (1987) Primary structure of rat cardiac beta-adrenergic and muscarinic cholinergic receptors obtained by automated DNA sequence analysis: further evidence for a multigene family. *Proc. Natl. Acad. Sci. USA.*, **84**: 8296.

6 Anderson, S. (1981) Shotgun DNA sequencing using cloned DNAse 1-generated fragments. *Nucleic Acids Res.*, **9(13)**: 3015-3027.

7 Gardner, R.C. Howarth, A.J., Hahn, P., Brown-Luedi, M., Shepherd, R.J., and Messing, J. (1981) The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.*, **9(12)**: 2871-2888.

8 Deininger, P.L. (1983) Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal. Biochem.*, **129**: 216-223.

9 Aach, J., Bulyk, M.L., Church, G.M., Comander, J., Derti, A., and Shendure, J. (2001) Computational comparison of two draft sequences of the human genome. *Nature,* **409**: 856-859.

10 Waterston, R.H., Lander, E.S., and Sulston, J.E. (2002) On the sequencing of the human genome. *PNAS*, **99(6)**: 3712-3716.

11 The International Human Genome Mapping Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**: 860-921.

12 Venter *et al.* (2001) The sequence of the human genome. *Science,* **291**: 1304-1351.

13 The Sanger Centre and The Washington University Genome Sequencing Center. (1998) Toward a complete human genome sequence. *Genome Res.*, **8**: 1097-1108.

14 Osoegawa, K. *et al.* (2001) A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.*, **11(3)**:483-496.

15 Shizuya, H., *et al.* (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA.*, **89**: 8794-8797.

16 The International Human Genome Mapping Consortium. (2001) A physical map of the human genome. *Nature,* **409**:934-941.

17 Weber, J.L., and Myers, E.W. (1997) Human whole-genome shotgun sequencing. *Genome Res.* **7**:401-409.

18 Green, P. (1997) Against a whole-genome shotgun. *Genome Res.*, **7**: 410-417.

19 Marshall, E., and Pennisi, E. (1998) Hubris and the human genome. *Science,* **280**: 994-995.

20 Collins F.S. (2001) Contemplating the end of the beginning. *Genome Res.*, **11**: 641-643.

21 Wadman, M. (1999) Human Genome Project aims to finish 'working draft' next year. *Nature,* **398**:177.

22 Pennisi, E. (2002) What's next for the Genome Centers? *Science online,* **291**(5507): 1204.

23 Pennisi, E. (2002) Genome Centers push for polished draft. *Science,* **296**: 1600-1601.

24 Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29(22)**: 4633-4642.

25 Korenburg, J.R. and Rykowski, M.C. (1988) Human genome organization: Alu, Lines, and the molecular structure of metaphase chromosome bands. *Cell,* **53**: 391-400.

26 Dunham, I., *et al.* (1999) The DNA sequence of human chromosome 22. *Nature,* **402**: 489-495.

27 Pennisi, E. (2002) Charting a genome's hills and valleys. *Science,* **296**: 1601-1603.

28 Caspersson, T., *et al.* (1970) Differential Binding of Alkylating Fluorochromes in human chromosomes. *Exptl. Cell Res.*, **60**: 315-319.

29 Korenburg, J.R., Therman, E., and Denniston, C. (1978) Hotspots and functional organization of human chromosomes. *Hum. Genet.*, **43**: 13-22.

30 Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228(4702)**: 953-958.

31 Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241(1)**: 3-17.

32 Zoubak, S., Clay, O., Bernardi, G. (1996) The gene distribution of the human genome. *Gene*, **174(1)**: 95-102.

33 Claverie, J-.M. (2000) From bioinformatics to computational biology. *Genome Res.,* **10**: 1277-1279.

34 Das, M., Burge, C.B., Park, E., Colinas, J., and Pelletier, J. (2001) Assessment of the total number of human transcription units. *Genomics,* **77(1-2)**: 71-78.

35 Claverie, J.-M. (2000) Do we need a huge new centre to annotate the human genome? *Nature,* **403**: 575.

36 Wright, F.A. *et al.* (2001) A draft annotation and overview of the human genome. *Genome Biology,* **2(7)**: RESEARCH0025. (online only)

37 Benson, D.D., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. (2000) GenBank. *Nucleic. Acids Res.*, **28(1)**: 15-18.

38 Boguski, M.S. and Schuler, G.D. (1995) ESTablishing a human transcript map. *Nature Genet.*, **10**: 369-371.

39 Wilcox, A.S., Khan, A.S., Hopkins, J.A., and Sikela, J.M. (1991) Use of 3` untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res.*, **19(8)**: 1837-1843.

40 Putney, S.D., Herlihy, W.C., and Schimmel, P. (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature,* **302**: 718-721.

41 Milner, R.J., and Sutcliffe, J.G. (1983) Gene expression in rat brain. *Nucleic Acids. Res.*, **11(16)**: 5497-5520.

42 Adams, M.D., *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and Human Genome Project. *Science,* **252**: 1651-1656.

43 Caron, H. et al. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science,* **291**: 1289-1292.

44 Ewing, B. and Green, P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.*, **25**: 232-234.

45 Martin, K.J., and Pardee, A.B. (2000) Identifying expressed genes. *Proc. Natl. Acad. Sci.* USA, **97(8)**: 3789-3791.

46 Quackenbush, J., Liang, F., Holt, I., Pertea, G. and Upton J. (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, **28(1)**: 141-145.

47 Deloukas, P., *et al.* (1998) A physical map of the 30,000 human genes. *Science,* **282**:744-746.

48 Roest-Crollius, H, *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetradon Nigroviridis* DNA sequence. *Nature Genet.*, **25**: 235-238.

49 Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J.. (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.*, **25**:239-240.

50 Rogic, S., Mackworth, A.K., and Ouellette, F.B.F. (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, **11**: 817-832.

51 Guigo, R. (1997) Computational gene identification. *J. Mol. Med.*, **75**: 389-393

52 Claverie, J-.M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, **6(10)**:1735-1744.

53 Altschul, S.R., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**: 403-410.

54 The National Center for Biotecnology information. http://www.ncbi.nlm.nih.gov/.

55 Lukashin, A.V., and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**:1107-1115.

56 Burge, C. (1997). Identification of complete gene structure in human genomic DNA. PhD thesis. Stanford University, Stanford, CA.

57 Royer-Pokora, B. *et al.* (1986) Cloning the gene for an inherited human disorder -chronic granulomatous disease- on the basis of its chromosomal location. *Nature,* **322**: 32-38.

58 Collins, F.S. (1995) Positional cloning moves from perditional to traditional. *Nature Genet.*, **9**: 347-350.

59 Schuler, G.D., *et al.* (1996) A gene map of the human genome. *Science,* **274**(5287): 540-564.

60 Harshman, K. *et al.* (1995) Comparison of the positional cloning methods used to isolate the BRCA1 gene. *Hum. Mol. Gen.*, **4(5)**: 1259-1266.

61 The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I &II Team. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature,* **420**:563-573.

62 Mouse Genome Sequencing Consortium. (2002) Initial sequence and comparative analysis of the mouse genome. *Nature*, **420**: 520-562.

63 Monaco, A.P., Neve, R.L., Colletti-Feener C., Bertelson, C.J., Kurnit, D.M., and Kunkel, L.M. (1986) Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature*, **323**: 646-650.

64 Cross, S.H., and Bird, A.P. (1995) CpG islands and genes. *Curr. Opin. Genet. Dev.*, **5(3)**: 309-314.

65 Cross, S.H., Clark, V.H. and Bird, A.P. (1999) Isolation of CpG islands from large genomic clones. *Nucleic Acids Res.*, **15(10)**: 2099-2107.

66 Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics,* **13(4)**: 1095-1107.

67 Cross, S.H., Charlton, J.A., Nan, X., and Bird, A.P. (1994) Purification of CpG islands using a methylated DNA binding column . *Nature Genet.*, **6**: 236-244.

68 Rommens, J.M., *et al.* (1989) Identification of the Cystic Fibrosis gene: chromosome walking and jumping. *Science,* **245**: 1059-1065.

69 Riordan, J.R., *et. al.* (1989) Identification of the Cystic Fibrosis Gene: cloning and characterization of complementary DNA. *Science,* **245**: 1066-1072.

70 Kerem, B.-S., *et. al.* (1989) Identification of the Cystic Fibrosis gene: genetic analysis. *Science,* **245**: 1073-1080.

71 Duyk, G.M., Kim, S., Myers, R.M., and Cox, D.R. (1990) Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Prot. Natl. Acad. Sci. USA*, **87**: 8995-8999.

72 Buckler, A.J., Chang, D.D., Graw, S.L., Brook, D., Haber, D.A., Sharp, P.A., and Housman, D.E. (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. USA*, **88**: 4005-4009.

73 Church, D.M., Stotler, C.J., Rutter, J.L., Murrell, J.R., Trofatter, J.A., and Buckler, A.J. (1994) Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.*, **6**: 98-105.

74 Das, M., Harvey, I, Chu, L.L., Sinha, M., and Pelletier, J. (2001) Full-length cDNAs: more than just reaching the ends. *Physiol. Genomics,* **6**:57-80.

75 Parimoo, A., Patanjali, S.R., Shukla, H, Chaplin, D.D., and Weissman, S.M. (1991) cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA.*, **88**: 9623-9627.

76 Tambini, C.E., George, A.M., Rommens, J.M., Tsui, L.-P., Scherer, S.W. and Thacher, J. (1997) The *XRCC2* DNA repair gene: identification of a positional candidate. *Genomics,* **41**:84-92.

77 Engelen, J., Hamers, A., Schrander-Stumpel, C., Mulder, H., and Poorthuis, B. (1992) Assignment of the aspartylglucosaminidase gene (AGA) to 4q33-q35 based on decreased activity in a girl with a 46,XX,del(4)(q33) karyotype. *Cytogenet. Cell Genet.*, **60**: 208-209.

78 Eberle, M.A., Pfutzer, R., Pogue-Geile, K.L., Bronner, M.P., Crispin, D., Kimmey, M.B., Duerr, R. H., Kruglyak, L., Whitcomb, D.C., and Brentnall, T.A. (2002) A new susceptibility locus for autosomal dominant pancreatic cancer maps to chromosome 4q32-34. *Am. J. Hum. Genet.*, **70**: 1044-1048.

79 Kaukonen, J., Zeviani, M., Comi, G.P., Piscaglia, M.-G., Peltonen, L., and Suomalainen, A. (1999) A third locus predisposing to multiple deletions of mtDNA in autosomal dominant progressive external ophthalmoplegia. (Letter) *Am. J. Hum. Genet.*, **65**: 256-261.

80 Rommens, J.M., Mar, L., McArthur, J., Tsui, L.-C., and Scherer, S.W. (1994) Towards a transcriptional map of the q21-q22 region of chromosome 7. In "*Identification of Transcribed Sequences*", (U. Hochgeschwender and K. Gardiner eds.), Plenum Press, New York. pp. 65-79.

81 Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989*) Molecular Cloning: A Laboratory Manual*, 2nd Edt., Cold Spring Harbor Laboratory Press: Plainveiw, New York.

82 Southern, E.M. (1975) Detection of Specific Sequences Amoung DNA Fragments Seperated by Gel Electrophoresis. *J. Mol. Biol.*, **98**: 503-517.

83 Inoue, H., Nojima, H., and Okayama, H. (1990) High efficiency transformation of *Echerichia coli* with plasmids. *Gene,* **96(1)**: 23-28.

84 Wheeler, D.L., *et. al.* (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28(1)**:10-14.

85 Baker, W., *et. al.* (2000) The EMBL Nucleotide sequence database. *Nucl. Acids Res.*, **28**:19-23.

86 Harger, C., *et. al.* (2000) The genome sequence database. *Nucleic Acids Res.*, **28(1)**:31-32.

87 Maglott, D.R., *et. al.* (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28(1)**:126-128.

88 Zhao, S. (2000) Human BAC Ends. *Nucleic Acids Res.*, **28(1)**:129-132.

89 Rodriguez-Tomé, P., and Lijnzaad, P. (2000) RHdb: the Radiation Hybrid databases. *Nucleic Acids Res.*, **28(1)**:146-147.

90 Kikuno, R., *et. al.* (2000) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.*, **28(1)**:331-332.

91 Wolfsberg, T.G., et al. (2002) A user's guide to the human genome. *Nature Genet.*, **32**.

92 Herzog, H. *et. al.* (1997) Overlapping gene structure of the human neuropeptide Y receptor subtypes Y1 and Y5 suggests coordinate transcriptional regulation. *Genomics,* **41**:315-319.

93 Gerald, C., *et al.* (1996) A receptor subtype involved in neuropeptide-Y-induced food intake. *Nature,* **382(6587)**: 168-171.

94 Thierry-Mieg, D.J., Thierry-Mieg, Y., Potdevin, M., Sienkiewicz, M., and Simonyan, V. Construction and automatic annotation of cDNA-supported genes using Acembly, www.humangenes.org: unpublished.

95 Online Mendelian Inheritance in Man, OMIM (TM). (2000) McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). World Wide Web URL:http://www.ncbi.nlm.nih.gov/omim/

96 Bennett, E.P., Hassan, H., Hollingsworth, M. A., and Clausen, H. (1999) A novel human UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase, GalNAc-T7, with specificity for partial GalNAc-glycosylated acceptor substrates. *FEBS Lett.*, **460**: 226-230.

97 Wanschura, S., Schoenmakers, E.F., Huysmans, C., Bartnitzke, S., Van de Ven, W.J., and Bullrdiek, J. (1996) Mapping of the human HMG2 gene to 4q31. *Genomics,* **15(2)**: 264-265.

98 Gabellini, D., Green, M.R., and Tupler, R. (2002) Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell,* **110**: 339-348.

99 Laherty, C.D. *et al.* (1998) SAP30, a component of the mSin3 co-repressor complex involved in N-CoR-mediated repression by specific transcription factors. *Molec. Cell.,* **2**: 33-42.

100 Dandoy-Dron, F. *et al.* (1998) Gene expression in scrapie: cloning of a new scrapie-responsive gene and the identification of increased levels of seven other mRNA transcripts. *J. Biol. Chem.,* **273**: 7691-7697.

101 Dron, M. *et al.* (1998) Characterization of the human analogue of a scrapie-responsive gene. *J. Biol. Chem.,* **273**:18015-18018.

102 Yamagishi, H., Olson, E.N., Srivastava, D. (2000) The basic helix-loop-helix transcription factor, dHAND, is required for vascular development. *J. Clin. Invest.,* **105**:261-270.

103 Cenciarelli, C., Chiaur, D.S., Guardavaccaro, D., Parks, W., Vidal, M., and Pagano, M. (1999) Identification of a family of human F-box proteins. *Curr. Biol.,* **9**:1177-1179,

104 Winston, J.T., Koepp, D.M., Zhu, C., Elledge, S.J., and Harper, J.W. (1999) A family of mammalian F-box proteins. *Curr. Biol.,* **9**:1180-1182.

105 Leroy, E., Boyer, R., Auburger, G., Leube, B., Ulm, G., Mezey, E., Harta, G., Brownstein, M.J., Jonnalagada, S., Chernova, T., Dehejia, A., Lavedan, C., Gasser, T., Steinbach, P.J., Wilkinson, K.D., and Polymeropoulos, M. H. (1998) The ubiquitin pathway in Parkinson's disease. (Letter) *Nature,* **395**: 451-452.

106 Lincoln, S., Vaughan, J., Wood, N., Baker, M., Adamson, J., Gwinn-Hardy, K., Lynch, T., Hardy, J., and Farrer, M. (1999) Low frequency of pathogenic mutations in the ubiquitin carboxy-terminal hydrolase gene in familial Parkinson's disease. *Neuroreport* , **10**: 427-429.

107 Liu, Y., Fallon, L., Lashuel, H.A., Liu, Z., and Lansbury, P.T., Jr. (2002) The UCH-L1 gene encodes two opposing enzymatic activities that affect alpha-synuclein degradation and Parkinson's disease susceptibility. *Cell,* **111**: 209-218.

108 Kikuno R., Nagase T., Waki M. *et al.,* (2002) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.,* **30**: 166-168.

109 Ohara, O., Nagase, T., Ishikawa, K.-I. *et al.* (1997) Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins. *DNA Res.,* **4**:53-59.

110 Yeh R.F., Lim L.P., and Burge C.B. (2001) Computational inference of homologous gene structures in the human genome. Genome Res. **11(5)**: 803-816.

111 Roux, A-F., Rommens, J.M., Read, L., Cuncan, A.M.V., and Cox, D. (1997) Physical and transcription map in the region 14q24.3: identification of six novel transcripts. *Genomics,* **43**:130-140.

Appendix 1. Sequences of oligonucleotides.

Map Element	Forward Primer	Sequence	Reverse Primer	Sequence	Product size bp
LOC256573	LOC206001-F1	GTCTGCCAAAACCAGGATTC	LOC206001-R1	AACAGATTTTGGCACCTTGC	342
LOC256573	LOC206001-F2	AAAGGGCTGTATCCCCAAGT	LOC206001-R2	GTGTTGGGATGCTCAGTGTG	288
GALNT7	GAL-FA	GCCAAAAATGAACAAGAGCAC	GAL-RA	TTCCCCATACAGGTGACAAG	2331
GALNT7	GAL-FB	GAGAGAAAGCCAAGCCATTG	GAL-RB	TGTGGCAGTGCTTCAAAAAG	
HMG2	HMG2-F	GGCAAAAGTGAAGCAGGAAA	HMG2-R	AAAAACGAAATGCAACATCCT	480
SAP30	SAP30A-F1	GTGCCGCTGTCTAACTTGGT	SAP30A-R1	TGGGAGATGCTCTTCTGGAT	398
SAP30	SAP30B-F1	CAGCACTGTCCACTGTTTCG	SAP30B-R1	CTTGCGCTCTTATCCAGCTC	498
LOC133123	LOC133123-F	TATGAAGGCCAAGTGGAGGT	LOC133123-R	TCGTACAGCAGCATGAGCTT	375
LOC166847	LOC166847-F1	TGGAAGACGCTTGGAACTTT	LOC166847-R1	CGGTGTTGAGGTGTGAGAAA	386
LOC152950	LOC152950-F	GGCCAAGTTTCACACGAAGA	LOC152950=R	TTAATTAACGTGCCGCGAAG	332
HAND2	HAND2A-F1	AGGGCGAAATGAGTCTGGTA	HAND2A-R1	GTCCATGAGGTAGGCGATGT	467
HAND2			HAND2-R	TCCTCTTTCACGTCGGTCTT	544
LOC152952	LOC152952-F	CACCCCACGATTAGCTGAGT	LOC152952-R	GTGTGGGGAACCGAGAAGAG	
LOC166476	LOC166476-F1	AGGCTGCACGCTTCTTTATG	LOC166476-R1	GCTGTTTCTTTGGCTTGTGA	501
LOC166476	LOC166476-F2	CAAGAGACAACCGGAGAAGG	LOC166476-R2	GACCCTCTCTGTCACCCTCA	420
LOC166475	LOC166475-F1	TGCATGTTCAGACTCCCAAG	LOC166475-R2	AGACATTGGACTCCCACAGC	
KIAA1712	KIAA1712-F	CCTTCTGCCTACCGAGTTCA	KIAA1712-R	ATAGACAGGGCAGCACAAGC	1202
KIAA1712	1712-FA	CTTCAGTTGGTTTCCTCC	1712-RA	GTAGCCATTCTAGCCACC	
KIAA1712	1712-FB	TGATAAAGGGAGACCCAGCA	1712-RB	TCAGGCTGGTCTCAAACTCC	
KIAA1712	KIAA1712-E1F	TTTCCTATTGTTGCCGGAAG	KIAA1712-E1R	CCGACCGTTTGTATTTGGAA	333
KIAA1712	KIAA1712-E2F	CCCTGGACTTACCCCTTTGT	KIAA1712-E2R	CAGCTACTTGGGAGGCTGAG	371
KIAA1712	KIAA1712-E2F2	TCCAGGGAAGGTGAGAGGTA	KIAA1712-E2R2	ACCCACCAGGACAAAAAGAA	356
KIAA1712	KIAA1712-E3F	TGACCAAAATAGAATCAAGAAGC	KIAA1712-E3R	CTCAGGTAGTCCCCCATGTT	290
KIAA1712	KIAA1712-E4F	TCCAACTTTGTGGAATTATGCTT	KIAA1712-E4R	AAAATATCCTGTCAATTGTATAT GCTG	326
KIAA1712	KIAA1712-E5F	TTACAAGAGTCTGTTTCCTCTGC	KIAA1712-E5R	TTGCTGTCTAGGGTTTTGATTTT	262
KIAA1712	KIAA1712-E6F	TCATTAGAGTCTCATTATTTGTC AGC	KIAA1712-E6R	TGATATTTTCTTTGTGGTTGCAT	273
KIAA1712	KIAA1712-E7F	TGGGAGGAAAACATTGAAGTG	KIAA1712-E7R	TTCCAAAGTTTAGAGGCTGAAA	300
KIAA1712	KIAA1712-E8F	AGATAACGCTGTGGAAGTGAG	KIAA1712-E8R	GGAAATGCTAAATATGAGAATAT TTGA	400
KIAA1712	KIAA1712-E9F	AGTAGCTTGCTATCTGCTTCCA	KIAA1712-E9R	GGAGCTTTGCAATTTCTAAACAA	383
KIAA1712	KIAA1712-E10F	TGCGGATAGCCATTTTTAAT	KIAA1712-E10R	AATGCTTGCTTGTTTTCCTCA	297
KIAA1712	KIAA1712-E11F	AATGAGGAAAACAAGCAAGCA	KIAA1712E-E11R	TTTCTGTTGCCCCTTTCTGT	273
Clone A04	F-A04	GTGGGGGCATTTATGGTTTT	R-A04	AAAGACATGCCCAAACAGACT	218
Clone 962	962-F	GCACATGAGGGCCAATTTAT	962-R	GTCTGCCAAAACCAGGATTC	155
Clone 965	F-965	AGCTATGCCCAATGTGGTTT	R-965	GGTGGAATGGAAAGGAACTCT	537
Clone 829	829-FB	TGTACCCTTCTTCCCCAACA	KAT02RB	TATTTGGGTGTGGGGAAGA	
Clone 829	829-FA	TCGAGATCTAGCGGCTATCA	KAT02RA	TGGAACACCACCTCTTCA	253
Clone 852	F-852	TTTCTCCCTTGGAACAGTGG	R-852	CCTGCTAGCCTCATCATTCG	301
Clone 858	858-F	CCCCAAAAGAAATTTTTCCA	858-R	CTCACCCTCCGTGAAAGTGT	185
Clone 955	KAT01FA	TTGCCTTGAAAGGCTGTAGAA	955-RB	ACGCACGTGTCCTACAAGGT	308
Clone 955			955-RA	CCCATGGATAGGGACAGAGA	207
Clone A32	A32-F	CTCTGTGCAAGTGCTGGAAA	A32-R	TGCAATTCCCACAAGAACAG	464
Clone 826	826-F	GGGAGGTGAATGTGTCATCA	KAT01RA	GGTTCAAAACTGGCATTTGAG	212
Clone A16	A16-F	GCAATCAATGCTGAGCTGAA	A-16R	TCGGTTGTTCCAAAATTTCAA	144
Clone A16			R-A16	TTCAGCTCAGCATTGATTGC	
Clone 944	F-944	GTTTCCAATCAGGGCAAGAA	R-944	GAATTCCAATTTTGGGGTCA	205
Clone A100	F-A100	TGCTGGGTTTAGCTCTCAGC	R-A100	TCCAGACATGATGGAGGAGA	207
RXG					

Appendix 2. The sequence in FASTA format of the 4q32 and 4q34 clones.

Bi-directional sequencing was undertaken for each clone utilizing the T3 and T7 primers within the pBluescript cloning vector. One direction has been shown for each clone, except when both directions are needed to obtain the full sequence. Base pairs are shown in lowercase if they were difficult to call. Base pair ambiguities were designated following the IUB ambiguity codes:

R = G or A	S = G or C (strong)	H = A, C or T
Y = T or C	W = A or T (weak)	V = A, C or G
K = G or T	B = C, G or T	N = A, C, T or G
M = A or C	D = A, G or T	

The clone number is in bold. Primer sequence (CGGAATTCTCGAGATCTA`**B**`**C**)` is underlined, refer to Figure 3.1. Internal *EcoR1* sites are shown in red (GAATTC). The second sequence represents the same clone with the repetitive elements masked by "N"s utilizing the RepeatMasker webserver.

4q32 clones.

>ktgallery04_11_Samp_**383**.
TCGAGATCYAGCGCAGGAGGTGATATCtAGAGGGAGAGAGAAACACATAGAGCTCAGACA
CTGGGAACTTGTCAGCTAGGTCAGGGATTTTGGTTTTTATCTTAAAAGTACGTGAAAATA
TGCTAAAGTTATTAATGAGTGCAATAACATGATCACTCAGGCTCTTGTGTGGAGAGTGGA
CTTAAGATGCATATATAGAAATGTGGAAGACTAGCGAGGATGTGGTTGGGGCAGTTTAGG
TGAGAAGTGTCTTGACAGCTTGCAATAGAGTGGGGACCGTCAATTTGGAGGGGAAGCAAA
GTTTGGGAATGTATTTAGAAGGTAGATATAAAAGAATTTAGTGATGAGTTCAATCTAAAG
GGTAAACAAGAGATATCAAAGATGTTTCCTTGGTTTCTGGTTTAAGTACACAGATGAATG
AAAATCTGGAAAAGTGTGGTGTTATGTGAAGGTGTGCATACTACGAGGGCTGTTTTTGAA
ACATGAAGGTGGAACTGTTAAGTAAGCAAGATATACCAATCTAGAGGTTGGGTGTGAAGC
AATATATATAATTAAGACCAGAGATCTCGA

>ktgallery04_11_Samp_**383**.
TCGAGATCYAGCGCAGGAGGTGATATCTAGAGGGAGAGAGAAACACATAG
AGCTCAGACACTGGGAACTTGTCAGCTAGGTCAGGGATTTTGGTTTTTAT
CTTAAAAGTACGTGAAAATATGCTAAAGTTATTAATGAGTGCAANNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNAATCTGGAAAAGTGTGGTGTTATGTGAA
GGTGTGCATACTACGAGGGCTGTTTTTGAAACATGAAGGTGGAACTGTTA
AGTAAGCAAGATATACCAATCTAGAGGTTGGGTGTGAAGCAATATATATA
ATTAAGACCAGAGATCTCGA

>ktgallery03_07_Samp_**5**.
GAATTCTCGAGATcTGTCCCCAGAGCAAAAAGCYGCCACTAGCtTAACCTCTCtATGCTT
TGCTTCTTTCATCCTAAAAGTATAagGTCCACTGCAGGGATTGTAAGAGATGAAATATGA
aAATTGTTTTACAAATgACTATCTCATACACAGCCCATAATACCACTGCTATTATTACCTT
GAAGTAGTGTTAAaaAGTCTGCaATTTCAGTTCTCACTAGCAAACAAAAAAAACCCAGGG
AATAGTTATGTTTGCCCTGCAAGTAAATACTGTTCTAGAAATACACAAATTTAAAGTATAG
TGGTTCTCAAAGAGTCTGGAGACCCTTAATGGTCCCTACAACCCTTTCTTTGAAAGGCTG
TGAAAAGTCAAACTATTTTCATTGTACTACTGAGATATTACATAGTTTTCAtTCtCCTGATCT
CACAAGTGCACAGTGGAGAAAAAaAAaAAaAAaaaKatCYcGRGAaTc

>ktgallery03_07_Samp_**5**.
GAATTCTCGAGATCTGTCCCCAGAGCAAAAAGCYGCCACTAGCTTAACCT
CTCTATGCTTTGCTTCTTTCATCCTAAAAGTATAAGGTCCACTGCAGGGA
TTGTAAGAGATGAAATATGAAAATTGTTTTACAAATGACTATCTCATACA
CAGCCCATAATACCACTGCTATTATTACCTTGAAGTAGTGTTAAAAAGTC
TGCAATTTCAGTTCTCACTAGCAAACAAAAAAAACCCAGGGAATAGTTAT
GTTTGCCCTGCAAGTAAATACTGTTCTAGAAATACACAAATTTAAAGTAT
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAAAAAA
AAAAAAAAAAKATCYCGRGAATC

>ktgallery06_04_Samp_**41**.
GGAATTCTCGAGATCTTTTTTTTTTTTTAAAAGCAAAATGTTTAAGAACAACATAATGCA
ACATAGAACACAAGAGAACAGAACTGAAAACCATCCTAGAGACTTAGTCCAACCCCTTCT
TTTAATATTTGAAGTAAATGAGAACAAGTGAAATTAATTTGAAAGAGTAGGtATACATTA
AGTGATTTTTATTATTCTTTACATAATTGAAAATATTTGAATTGACTGtGAAGTGAGTCA
AAAAATTAAGTTGCAGTTATCACAAAAAGTGAAGTTAACAGTATGCtCATGATAAAAAGT
GGTTCTAGAATACAGCTGTCCTTATTCCTAGCATAGTATTTTAAAAAATACTACAGAATA
CCTGACTATAACCTCCtCACTATGGAAAGAAAAAACACACATGTACAAGTGCTTAGTGTC
TAAGAAATTTTTTAACAGCTAGATCTCGaGAATTC

>ktgallery06_04_Samp_**41**.
GGAATTCTCGAGATCTTTTTTTTTTTTTAAAAGCAAAATGTTTAAGAACA
ACATAATGCAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNTTTGAAAGAGTAGGTATACATTAAGTGATTTTTATTATTCTTT
ACATAATTGAAAATATTTGAATTGACTGTGAAGTGAGTCAAAAAATTAAG
TTGCAGTTATCACAAAAAGTGAAGTTAACAGTATGCTCATGATAAAAAGT
GGTTCTAGAATACAGCTGTCCTTATTCCTAGCATAGTATTTTAAAAAATA
CTACAGAATACCTGACTATAACCTCCTCACTATGGAAAGAAAAAACACAC
ATGTACAAGTGCTTAGTGTCTAAGAAATTTTTTAACAGCTAGATCTCGAG
AATTC

>ktgallery06_12_Samp_**01**.
GAATTCTCGAGATcTAGCGAGGATGTGGTTGGGGCAGTTTAGGTGAGAAGTGTCTTGACA
GCTTGCAATAGAGTGGGGACCGTCAATTTGGAGGGGAAGCAAAGTTTGGGAATGTATTTA
GAAGGTAGATATAAAAGAATTTAGTGATGAGTTTAATcTAAAGGGTAAACAAGAGATATc
AAAGATGTTYCTTGGTTTCTGGTTTAAGTAcacAGATGAATGAAAAtCTGGAAAAGTGTG
GTGTTATGTGAAGGTGTgCATACTACGAGGGCTGTTTTTGAAACATGAAGGTGGAMCTGT
TAAGTAASCAAGATATaCCAAtCTAGAGGTTGGGTGTGAASCAATATATATAATTAAGAY
CYCGAGAATYC

>ktgallery06_12_Samp_**01**.
GAATTCTCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNAATCTGGAAAAGTGTGGTGTTATGTG
AAGGTGTGCATACTACGAGGGCTGTTTTTGAAACATGAAGGTGGAMCTGT
TAAGTAASCAAGATATACCAATCTAGAGGTTGGGTGTGAASCAATATATA
TAATTAAGAYCYCGAGAATYC

4q34 Clones.

>gallery833_846_Samp_**833**.
GAATTCTCGAGATCTGCACACAGTCCACCTCTTCAGGATAATTTAGCAAGCGGASACCTG
CTCTAGGTTCCGTAAGCTTCTTTTTAAGTCACCTGTTGCCATTACTTAACATCAGATTCA
GCTCCAATTCTTGAAATCGTCAAATTTTTCACCTAGCAGAGCACAGACCAGCCAATCCCA
TCTGCCTAATAAAGAAAAAAAAAAAAAAGATCTCGAGAATTC

>A1-A23gallery_Samp_**04**.
GaaTTcTCGAGATCTAGCAGAGAGGTATAAATAAATCTGGAATGATGTGGGGGCATTTAT
GGTTTTTTAATAAATTAGGATTTAGGAATTTATCAACTGTGGGTGGTTTTGGTTGGCCTG
AATGGTGACACTGGCAGAATATTTTTGTGAACTGTTCTGCTTTAGTGTTGGATATGGTAC
TACTGGAGTCCTTATAAAGTGAAACACTGTCTGAACTCAGTTTGAAAGAAGTATCTGGCT
AATAGTCTGTTTGGGCATGTCTTTATTTTATAAAAATTCTGTAAAAAAAAAAAAAAAAAA
AAAAgatCtCGaGAaTTC

>gallery957_980_917_Samp_**962**.
GAATTCTCGAGATCTAGCACATGAGGGCCAATTTATGATTATGAATTTCTTAGATGATGG
AGTGAATACGTGTAGAGGGAGATTTCTCTTCAAAAGTTGAGGGAATGAGTTAATTTTAGT
TCTCTCTCACCTTGGGGATACAGCCCTTTTGGAATCCTGGTTTTGGCAGACCTATCTGCT
TTTCCCAAAAAAAAAGAAAAAAAAAAAAAAAAAA

>gallery957_980_917_Samp_**965**.
GAATTCTCGAGATCTTGCTTATCCAGCTATGCCCAATGTGGTTTAGCCAAACCTCTCATT
TTACATATGGCAGCATGGGTTATCTGGCCTTTTCACTGACATAGTTTCTTTCTATTCCAC
CCCTGTTGTGTCTAAGTTATGCTCTAAGAGATAATCAGTATTAACAAGAGAGAGGCTAGT
TTGTCTTATAGGTTCCCACAATAAAGAATGTCAGGTTCTAcgGCATAGCTAGATAATCAG
GATCTCTGGAAGGATGTGATAACAGATTGATAACTGTCTGAAACTAGTCTGCTTGCTGGT
GGAGAGGTTTATGTGCACAGACTGTAAGACGACTTACAGTAATACCTTGATAAATACCAG
TTTTGAAGACTTTCAGGTCATTTAACTAACCACACTGCTGAGAACCTGAGTCTTTCTTCA
GTGCATTTCATACCCTAGTTTTAAGAAATTTGAGCATTTTAAAGGAAATTAGAAATTTGA
GTCTAGCCAAATCTCTTAATCTGCATATAGCAGGGCAGGTTCTCTAGCCCTTTCACTGAC
AGAGTTCCTTTCCATTCCACCGCTAGATCTCGAGAATTC

>gallery825_832_Samp_**829**.
GAATTCTCGAGATCTAGCGGCTATCAGTTTATTATCTAGGAATTACAGTACTCATTTTGG
TCAGGTAAAATGTAGAATACTGTACCCTTCTTCCCCAACACCCAAATAAAGGCAAAACCA
TCTACCTATAAAATGCTTCTGCAAATCATTTAGATATTATCATATTTTAAGAGCTATTAA
AGAATGCTTCTTAAATAAGAAGACTACCGAATTATAGAAGTCCAAAAACAGTACCAAATT
ATGAAGAGGTGGTGTTCCACACT

>gallery825_832_Samp_**829**.
GAATTCTCGAGATCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

>ktgallery937_A37_Samp_**938**.
GAATTCTCGAGATCTTTTTTTTTTTTTTTTTTTGAtATTAAATATTTTTCAAAGAAACT
AAGGTACATCCATTTATGTCTCAcAcATAGGCCTGAACATTTCGGAAATGTTTGGCGAATTC

>ktgallery937_A37_Samp_**938**.
GAGGTCGACGGTATCGATAAGCTTGATATCGAATTCGCCAAACATTTCCGAAATGTTCAG
GCCTATGTGTGAGACATAAATGGATGTACCTTAGTTTCTTTGAAAAATATTTAATATCAA
AAAAAAAAAAAAAAAAAAARATCYCG

>ktgallery852_A100_Samp_**852**.
GAATTCTCGaGATCTAGCTGTTGTGTCTGTGTGTGTGTGTGTGCGTGTATTAAGAGTA
ACCTGCTAGCCTCATCATTCGGGGCAACACGCTGTGTGTGTGTGTGTGTGCGTGTGTG
TGTGTGTACTATGAGTAACCTGCTGCCTCGTCATTCAGAGGCAACACGCTGTGTGCCCAC
ATGGCAAGTTTACTTTTTCTAGCTACCTGCAGTCAATAAAATATACAACTGACATTTTTT
ATCCTTAAAATAATCATTCTTACTTTTAGCACCCACTGGAACCCAAAAGAAACTGACCAA
TGAATCACTCAAATTCCAGGCCTCATAATCAGAGAAAGGAATCCACTGTTCCAAGGGAGA
AAAAAATAGATTGTAGTTATTTTTAAAAAGAGTTAAAGGAATAGTAGGAAAAGTATCTCA
ATTCACCAGAAAACAAAAATACTGAACAAATGTATATTACTTA

>ktgallery858_A52_Samp_**858**.
GAATTCTCGAGATCTAGTATGGTAGTAATTAAAACCCTAAATTATACTAGCTTCTTTTCC
CCAAAAGGAATTTTTCCACTTTTTTTTTTTTTTTGAGATTCATTGATCTTCTTCACTCCC
AGTAGTAAGTTAAGCATCTAAGAGACACTGTCCTTTCAAGATTCCACATGCCACCCAGTG
CACCATTTACTCTGTACATAACATGGGCTGTCCATGCTGCTACACACTTTCACGGAGgGT
GAGGAGAATTC

>gallery825_832_Samp_**832**.
GAATTCAAATTTTATACGTGGAGAGGATAATATTAACCACTGAACACAAACAGGAAGACT
GAGGGAGAAAGGCTCAGGCAGGTGCATTCTGATGGTTTCCCAATAATTTACTTCTACAGC
CTTTCAAGGCAACGATGATAATTCAAGATCTCGAGAATTC

>ktgallery937_A37_Samp_**943**.
GAATTCTCGAGATCTTGAATTATCATCGTTGCCTTGAAAGGCTGTAGAAGTAAATTATTG
GGAAACCATCAGAATGCACCTGCCTGAGCCTTTCTCCCTCAGTCTTCCTGTTTGTGTTCA
GTGGTTAATATTATCCTCTCCACGTATAAAATTTGAATTC

>ktgallery937_A37_Samp_**954**.
GAATTCTCGAGATCTTGAATTATCATCGTTGCCTTGAAAGGTTGTAGAAGTAAATTATTG
GGAAACCATCAGAATGCACCTGCCTGAGCCTTTCTCCCTCAGTCTTCCTGTTTGTGTTCA
GTGGTTAATATTATCCTCTCCACGTATAAAATTTGAATTC

>gallery911_955b_Samp_**955**.
GAATTCTCGAGATCTTGAATTATCATCGTTGCCTTGAAAGGCTGTAGAAGTAAATTATTG
GGAAACCATCAGAATGCACCTGCCTGAGCCTTTCTCCCTCAGTCTTCCTGTTTGTGTTCA
GTGGTTAATATTATCCTCTCCACGTATAAAATTTGAATTC

>gallery833_846_Samp_**835**.
GAATTCAGCAGGACCTTGTAGGACACGTGCGTTTCATTTTAAGCTACATCCCATTCTCAG
TTCTCTGTCCCTATCCATGGGTGGGAAGGTTCAACCCTGTTCCATGGACATAGCTGATGG
GACAGGAGTATCCATAGCTAGATCTCGAGAATTC

>ktgallery937_A37_Samp_**937**.
GAATTCAAATTTTATACGTGGAGAGGATAATATTAACCACTGAACACAAACAGGAAGACT
GAGGGAGAAAGGCTCAGGCAGGTGCATTCTGATGGTTTCCCAATAATTTACTTCTACAGC
CTTTCAAGGCAACGATGATAATTCAAGATCTCGAGAATTC

>ktgallery937_A37_Samp_**939**.
GAATTCTCGAGATCTTGAATTATCATCGTTGCCTTGAAAGGCTGTAGAAGAAAATTATTG
GGAAACCATCAGAATGCACCTGCCTGAGCCTTTCTCCCTCAGTCTTCCTGTTTGTGTTCA
GTGGTTAATATTATCCTCTCCACGTATAAAATTTGAATTC

>gallery833_846_Samp_**845**.
GaTTCTCGAGATCTAGCTTGGRTACTCCTGTCCCATCAGCTATGTCCATGGaACAgGGKT
GRACCTTCCCACCCATGGATAGGGRCAGAGAACTGAGAGTGGGATGTAGCTTAAAATGAA
ACGCATGTGTCTTACAAGGTCCTGCTGAATTC

>A29_A45gallery_Samp_**A32**.
GAATTCTCGAGATCTTTTTTTTTTTCTTTTTTTCTTTTTTTTTTTTTTTTTTTTTTTCTTTT

>A29_A45gallery_Samp_**A32**.
GAATTCTCGAGATCTGCAGAGCGTCACTATTTTGCTTAAGAAAAAGAATCTCAGTGATGA
GTCTGGAAGTCTGAGGTTACTTTCCAGCACTTTAAAAAAGAAAA

>ktgallery852_A100_Samp_**A76**.
GAAtTCtCGAGAtCTTTTTttTTtttCCWtTTTTtHTTTcTCTtTTGTMAcCTTTTGcAA
CATGACAATTTTATGYCAAGAGAATGAAAAtGtCAGGtATtTCCATtACAGAGAGYCACA
TTTTCCCTTGATTGTTTcAGGCAGGAAGAAAGCGACTCAGATTAATCTGTTTAGAACTAT
TTtCAGTTATCTGTATATGTCATACATGCTGTCATGTTCTAAGACCACTTTGTTCtTGGA
TACtGCTTTCCAAAATCTATGGGATTctACtCTTAAAAAAAAAAAAAgAATATCTCGCTTT
ACATaTCATTTTAtTCTGGGGAMGCCTAaTtAGCaCATTAWTaTCAGAGAGCTGACAACC
tCCTACAtGTTATTYtAGGATTCcAtGTGCAAGACCATATtAtYCTAGTTATTTgAATGG
CAcGTATTCTcCACGAGCTAGATYtcGAGAATTC

>gallery825_832_Samp_**826**.
GAATTCTCGAGATCTCGTTCAGTGATAAATCTCTCTATTTTAAACTTAGGTTCAAAACTG
GCATTTGAGAAATGCTACTCTAATTCTTTTCTCATTCAGAATATTACTTTCATCTCACTG
AAATAAAAAGCAGCACATAACACAATCTACAATGTAAGATTTCAGGATGATAGGATCTTT
GTACTACAAAAAATGGAAATGGCTGTATCCTGCAAAATATATTTGCTGCAACATTCTTTA
TGATGACACATTCACCTCCCACAGATCTCGAGAATTC

>A7-A24gallery_Samp_**A16**.
GAATTCTCGAGATCTAGGAATAAAAGTCAACATGAATAAAATGGAGAAGCAATCAATGCT
GAGCTGAATTTTTCAAATAGAGTAGAgCTATTAATCATATTGGGTCAGAGATGTAATTCT
TTGGTGAGAGGCTAATAAATGAAATAAAATTATAGGTCTTTTTGAGACTAATTGAAATTT
TGGAACAACCGATAGAATAAAACTACATTCTAAAATTGTAGTTTTCAGATTCAGACTTAG
AACATTCCTTAATTAACTTAGCATTTGATTTATAGAGAACTTTAATATTTCAATAGTATT
ATCTATCAAATAGTTAATGAACTATTTTAACTTTCTTTGAGAAAGTTAAAAGGCAGATCT
CGAGAATTC

>ktgallery852_A100_Samp_**A18**.
GAATGTTAAAAYTgGAgATAtCtGCcTTTTAACTTCCTCAAAGAAAGTTAAAATAGTTCA
TTAACTATTTGATAGaTAATACTATTGAaATATTAAAGTTCTCTATAAATCAAATGCTAA
GTTAATTAAGGAATGTTCTAAGTCTGAATCTGGRAACTACAATTTTAGAATGTAGTTTTA
TTCTATCGGTTGTTCCATAATTTCAATTAGTCTCAAAAAGACCTATAATTTTATTTCATT
TATTAGTCTCTCACCAAAGAATTACATCTCTGACCTAATATGATTAATAGCTCTACTCTA
TTTGAaAAATTCAGCTCAGCATTGATTGCTTCTCCATTTTATTCATGTTGACTTTTATTC
CTAGGCTAGAGAATAACTCCATAAGGGAAGAAACTCTTCCTAAAAAAAAAAAAAAAAAAA
GATCTCGAGAATTC

>gallery925_987_Samp_**1040**.
GAATTCTCGAGATCTCAAAGTTTTCACCACCCCAAATCTGGAGACCTGGATCATAGAGAC
CAATTCAAAGAAGAACTCTCGTTCAATGGCAAATAATCCCCCAGCCGTGGCTGGGACCTG
TCCTTAGATATGGGAGCTACAAGTGGTGCATACCAGTTAACTGCCACCTCACAGTGGGCA
TCAAGGTATATCAAAACCTGTCCAAGTTTAGCCTTCTGAGCACCAATACTTCGTGCTTGA
ATTAAACCTTCCCTTCTTTCATTTCGAAATACCTTCACTAGGCCATTCCACAGCTTAATA
TATTCATCCAGTTTTTCTTATAAGTGTTCTTTATTACTGAAATCGTCAATTAACACAATT
TCTGCTAGATCTCGAGAATTC

>gallery911_955_Samp_**944**.
gaattctCGAGATCcAGCcGCAGGGYAAGTTGAGTTTCCAATCAGGGCAAGAAGGGTTTC
TGACCTATAAAAACAAtGGTTAGCTGTAAGGCATATATaAaCTTTTATTAACCTTGGACC
TGACTCAAGTAAAAATTATTACACTTATATTTACTATTTAATTTATAGAAGTGGCAAACT
GGTGACCTTTGGGCAGAATCTTACCTATAGACATGTTTTGACCCCAAAATTGGaATTC

>ktgallery1041_A70_Samp_**A100**.
GAATTCTCGAGATCTGCAGTTCATAAGTTTATTATCTATATCTGAAAGAATCATAGAAAA
TTGCTGGGTTTAGCTCTCAGCAGCCCGCTCCTGAGCTCTGAGGAAGCTTGCCTTCTTTTG
AGCTACCCGATCCTTCTTCTGAGCAAGGGACATTTTGGGACGGTTCCACCTCTTCTTTTT
AACTTCTTTCTTGGGCTTCTTTTCATAGACTGGATTCTCTCGTATAGCAGCATGAGCTTT
CTTATACATCTCCTCCATCATGTCTGGAGTTACGCTGTTCTTTAAAAAAAAAAAAAAAAA
AAGATCTCGAGAATTC