Extracting Knowledge from Gene Expression Data: A Case Study of Batten Disease

Simon M. Lin ¶
Duke Bioinformatics Shared Resource
Duke University Medical Center
Durham, NC 27710

Sumeer Dhar
Departments of Pediatrics and Neurobiology
Duke University Medical Center
Durham, NC 27710

Rose-Mary N. Boustany
Departments of Pediatrics and Neurobiology
Duke University Medical Center
Durham, NC 27710

ABSTRACT

Batten disease is a group of neurodegenerative disorders characterized by massive neuronal death. The devastating syndrome of Batten disease includes decline in both cognitive and motor skills, seizures, and loss of vision. The mechanism of this disease is not clear. Large-scale gene expression studies provide the possibility of identifying the molecular mechanism of this disease, and thus a first step to finding a cure. We proposed a prototype knowledge discovery and data mining (KDD-g) system to enable scientists to analyze the massive microarray data set creatively, form hypotheses intuitively, and draw insights directly into underlying mechanisms of diseases. The findings from this study establish that at the gene expression level, CLN9 is a unique form of Batten disease. It does not show any similarity with any of the other NCL variants. However, CLN9 share many common disease mechanisms with other types of Batten disease, such as abnormalities in the apoptosis pathway and protein degradation. Based on the results acquired in the preliminary data analysis using KDD, we conclude that this approach could be utilized in the future to study various clinical diseases.

Keywords

Batten disease, knowledge discover and data mining (KDD), microarray, gene expression, functional genomics, molecular mechanism.

1. INTRODUCTION

Neuronal ceroid lipofuscinoses (NCLs), collectively referred to as Batten disease, are the most common hereditary neurodegenerative diseases of childhood. The overall frequency is 1 in 12,500 births [1]. The clinical symptoms of Batten diseases include progressive motor and mental deterioration, seizures, loss of vision, and premature death. At present, there is no effective therapy. Previous studies have identified 8 types of NCL (CLN1 to CLN8). A new form of NCL, CLN9 (not fully classified yet), has been recently identified in our lab. Although the clinical picture categorized CLN as Juvenile form of Batten disease, enzyme assays and DNA mutation analysis have ruled out the possibility of these variants to be CLN1, CLN2 or CLN3.

Our goal is to identify a set of genes implicated in CLN9 and try to elucidate the pathways that cause this defect. With the advent of cDNA microarrays and DNA chips, genome-wide expression profiling has become available to identify genes whose expressions are altered in diseased states. [2]; [3]; [4]. In Affymetrix oligonucleotide arrays approximately 7000 genes are simultaneously analyzed in a parallel fashion. Analyzing and interpreting this genomic-scale information would aid us in pinpointing the changes in the expression levels of CLN9 deficient cell lines which would eventually help us in potential diagnostics and therapeutic applications.

The huge amount of data from microarray study of poorly understood phenomena, poses great challenges to traditional data analysis methods that are designed to confirm certain hypotheses. Knowledge discovery in databases (KDD) has emerged as a new technique to discover interesting, implicit, and previously unknown knowledge from large databases [5]. We proposed a prototype knowledge discovery and data mining system (KDD-g) to enable scientists to analyze the massive microarray data set creatively, form hypotheses intuitively, and draw insights directly into underlying mechanisms of diseases.

In this study, we used Affymetrix chips to study the differences in gene expression of fibroblasts from normal and Batten diseases patients (CLN1, CLN2, CLN3, CLN6 and CLN9), and demonstrated how data mining and knowledge discovery techniques may change the way scientists can analyze functional genomic data generated in the laboratory.

The rest of this communication is organized as follows: section 2 describes the material and methods; in section 3, the architecture of a KDD system specially designed for genomics, KDD-g, is shown; section 4, 5, 6, and 7 discuss four initial results that demonstrate the potential usefulness of applying KDD techniques in studying the mechanism of Batten disease; future studies are discussed in section 8.

¶ Correspondence to: Simon Lin, M.D., Duke Bioinformatics Shared Resource, Box 3958, DUMC, Durham, NC 27710. Tel: +1 (919) 681-9646. FAX: +1 (919) 681-8028. E-mail: lin00025@mc.duke.edu.

2. MATERIALS AND METHODS

The fibroblast cell lines used in the study consisted of a normal control cell line (ST99-20), which was derived from skin biopsy and the diagnosis was confirmed by brain biopsy and evaluation at Tufts New England Medical center by Dr. Joseph Elroy. The CLN1 deficient cell line was a kind gift from Dr. Hoffman at Texas South Western School. The diagnosis was confirmed by TPP and DNA analysis. The CLN2 cell line was derived from the skin biopsy done by Dr. Boustany at Duke University Medical Center and evaluated by DNA analysis and enzyme assay by Joseph Alroy at Tuft, and TPP evaluation was done in Dr. Peter Lobel's laboratory. The CLN3 deficient cell line was confirmed by DNA study and diagnosis was performed by Dr. Cathryn Simms at Mass General Hospital, MA. The CLN6 deficient cell line was derived from the skin biopsy done by Dr. Boustany in Costa Rica and Dr. Joseph Elroy performed the EM, and DNA study. CLN9 (1) deficient cell line was derived from skin biopsy and the diagnosis was confirmed by brain biopsy which was evaluated by Dr. Josephs Elroy at Tufts New England Medical Center. The CLN9 (2) deficient cell line was a kind gift from Dr. Alfred Kohlschultter, Germany. The cell lines were grown in culture medium DMEM (GIBCO BRL), supplemented with 10% fetal Bovine serum (Hyclone), glutamine, 1% penicillin-streptomycin (Hyclone). The cells were subcultured twice every week.

The DNA chips used in this study were Affymetrix GeneChip HuGeneFL, which consists of probes for a total of 7000 human genes. The preparation of total RNA, fluorescent labeling, and hybridization were performed as described in the Affymetrix technical manual. The image analysis was done by Affymetrix software. The raw measurements of the expression of each gene were taken from the Affymetrix software (AvgDiff). After data transformation and normalization [6], fold changes were computed. Genes with negative values were reset to a minimum value of 20 to eliminate infinite-fold changes. To determine which genes differed in the CLN9 vs. normal samples, a conservative pairwise 2-fold change criterion [7] was applied. Only genes with consistent Absent/Present calls (AbsCall from Affymetrix software) from all replicates, and with consistent greater than 2-fold change in all possible pairs of comparison of CLN9 vs. normal samples, were included for further discussion. Expression levels of the genes were visualized by an established color code [8].

3. KDD-g SYSTEM ARCHITECTURE

Expression profiling provides a rapid genome-wide survey of cellular states. It has been applied to further our understanding of various biological processes: from cell-cycle regulated genes in yeast, to the identification of genes involved in complex human diseases such as cancer [9]. The discovery of knowledge in microarray databases is becoming increasingly important, as the scientific investigators are being overwhelmed with an influx of data from genomic-scale studies [10].

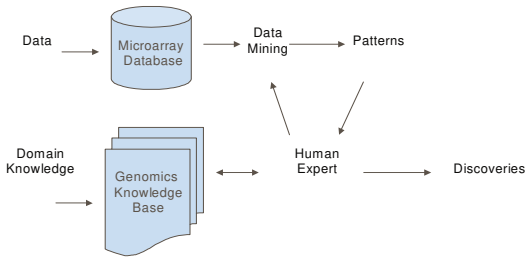

Fig. 1. System architecture of KDD-g.

Figure 1 shows the architecture of KDD-g, a prototype KDD system for functional genomics. Domain knowledge, such as functional genomic concept hierarchies, is stored in a knowledge base. Rules and patterns are discovered by the data-mining modules that include database interaction, feature selection, pattern extraction, and evaluations. New statistical, machine learning, and visualization algorithms can be added in this data-mining module.

While human insight into the complex disease mechanisms cannot be automated at this time, the perusal of the massive microarray dataset can be refined by the data mining tools to manageable patterns of interest. Thus, by appropriate use of domain knowledge in conjugation with the empirical application of discovery algorithms, the human expert is able to focus on likely disease targets. In this manner, new domain knowledge is acquired and fed back into the knowledge base for further refinement.

In the next four sections, we will demonstrate the usefulness of this system in the elucidation of the mechanisms of Batten disease.

4. CLN9 IS A NEW TYPE OF BATTEN DISEASE

Global gene expression profiles were compared in human fibroblast cell lines from normal adult individuals and patients with Batten disease. Figure 2 shows the patten of gene expression in CLN9 (d85, d86, d379, d380), CLN1(d382), CLN3(d381), CLN6(d88), and normal fibroblasts(d378, d844). The visualization of the gene expression matrix is facilitated by color-coding. A Self-organizing map [11] was used to order the genes.

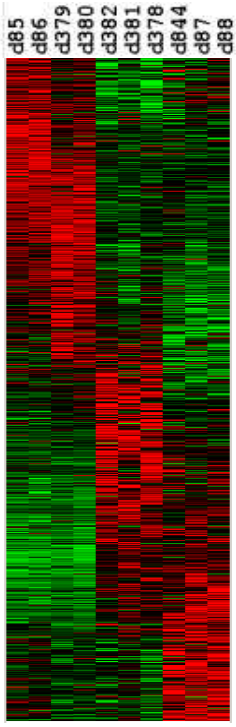

Fig. 2. Gene expression profile in fibroblast cell lines from normal adult and Batten disease patients. Colors represent relative levels of gene expression, with the brightest red indicating the highest level of gene expression, black indicating medium level of expression, and green depicting low level of expression. Each horizontal line represents a gene; each column represents a cell line. 2543 genes from the Affymetrix chip HuFL6800 are shown in this figure.

The findings from this study establish that at the gene expression level, CLN9 is a unique form of Batten disease. It does not show any similarity with any of the other NCL variants.

5. APOPTOSIS AS A MECHANISM OF BATTEN DISEASE

Neurodegenerative processes are generally characterized by the long-lasting course of neuronal death. Apoptosis has been proposed as the major mechanism of neuronal death in Batten diseases [12]. Apoptosis is the cell's decision to commit suicide, and therefore is referred to as programmed cell death. Activation of apoptosis in an individual cell is based on its own internal metabolism, external stimuli, environment, developmental, and genetic background. Such a situation occurs in most of the neurodegenerative disorders such as Alzheimer's, Parkinson's and Huntington's diseases. In these pathological situations, specific neurons undergo apoptotic cell death characterized by DNA fragmentation, increased levels of pro-apoptotic genes and apoptotic proteins. In this study, we observed the involvement of the apoptotic pathway in the disease mechanism of CLN9.

The myristoylated alanine-rich C kinase substrate (MARCKS) protein family has two known members, MARCKS itself and MARCKS-related protein (MRP, also called MacMARCKS or F52). They are essential for brain development and are believed to regulate the structure of the actin cytoskeleton at the plasma membrane. MacMARCKS gene was upregulated more than 5 folds in CLN9.

Caspases are a family of cysteine proteases related to interleukin-1 converting enzyme (ICE) and represent the effector arm of the cell death pathway. Proteases of the caspase family, especially caspase-1 (ICE)(-like), caspase-3 (CPP32/Yama/apopain)(-like) and caspase-8 (MACH/ FLICE/Mch5) proteases, are implicated in Fas (APO-1/CD95)-mediated apoptosis. Caspase-4 has recently been reported to be involved in Fas-mediated apoptosis [13]. Caspase 4 (U28014) was down-regulated more than 10 fold in CLN9, but at normal level in CLN 1, 2, 3, and 6.

BNIP2 (U15173) is a member of the BCL2/adenovirus E1B 19 kd-interacting protein (BNIP) family. It interacts with the E1B 19 kDa protein which is responsible for the protection of virally-induced cell death, as well as E1B 19 kDa-like sequences of BCL2 which is also an apoptotic protector. E1B 19 kDa and the Bcl-2 proteins, promote cell survival through interaction with a common set of cellular proteins. BNIP2 is down-regulated in CLN9 which, therefore, fails to provide necessary protection to the cells from undergoing apoptosis.

Chromosome segregation 1-like (CAS yeast homologue, Affymetrix accession # U33286) has been noted to have dual functions in mammalian cells, it is implicated both in apoptosis and as well as cell proliferation [14]. CAS is highly expressed in human tumor cell lines and in tissues that contain actively dividing cells. CAS expression has been shown to increase when resting human fibroblasts are induced to proliferate and decrease when they are growth-arrested. Thus, CAS appears to play an important role in both toxin and tumor necrosis factor-mediated cell death, as well as in cell proliferation. CAS is upregulated in all CLN9-deficient patient cell lines.

6. PERTUBATION OF PROTEIN DEGRADATION PATHWAY

Protein degradation plays important roles in cellular physiology: recycling amino acids; degrading regulatory proteins in response to changes in the environment; and cleaning up mutated or misfolded proteins. Protein degradation occurs in two locations: lysosome (Fig. 3) and cytosol (Fig. 4). In the present study, using DNA chips, we discovered the coordinated upregulation of cytosolic protein degradation in CLN9 patients, and down regulation of lysosomal protein degradation. Here we discuss nine genes in detail.

6.1 Lysosomal Protein Degradation Pathway

Batten disease has been characterized as a lysosomal storage disease. Several subtypes of Batten disease have been related to the enzymatic defect in the lysosome. The role of cathepsins in Batten disease was hypothesized by [1]). Recent study showed the defect of cathepsin D in congenital ovine NCL [15]. Here we showed PPGB (M22960), Cathepsin B (HG417-HT417) and L (X12451) was down-regulated in CLN9 patients. PPGB encodes a glycoprotein

that associates with lysosomal enzymes, and has cathepsin A-like activities. Cathepsins are lysosomal cysteine (thiol) proteinases that play a role in intracellular protein catabolism. Pharmacology inhibition of cathepsin B and L by *N*-CBZ-L-phenylalanyl-L-alanine-diazomethylketone produces an accumulation of lysosome-related dense bodies [16].

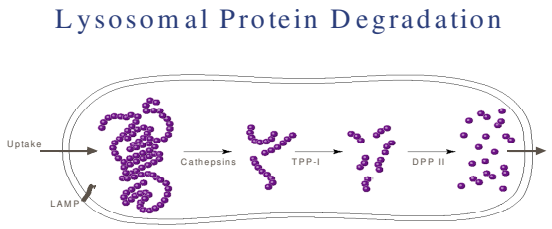

Fig. 3. Lysosomal protein degradation pathway.

Lysosome associated membrane glycoprotein was identified as a receptor on the lysosome membrane to selectively uptake proteins for lysosomal degradation [17]. It was an obvious candidate for some variants of Batten disease although at present no mutations in LAMP family proteins was shown [1]. In the present study, we revealed LAMP-3 was down regulated in CLN9 patients.

6.2 Cytosolic Protein Degradation Pathway

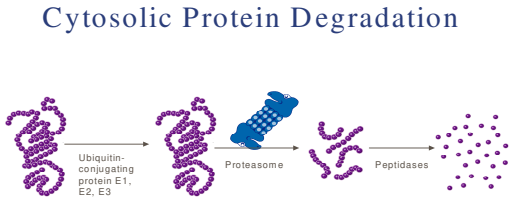

Fig. 4. The Ubiquitin-proteasome pathway in the cytosol.

In the cytosol, proteasome is responsible for initiating the main part of protein degradation [18]. This system is involved in a number of regulatory processes, including control of cell cycle, stress response, and apoptosis. The components in the cytosolic protein degradation system are shown in Fig. 4. In contrast to the lysosomal pathway, this cytosolic protein degradation pathway was generally upregulated in CLN9.

Ubiquitin plays a central role in cytosolic protein degradation. A variety of neuropathologic conditions including Parkinson disease, Neimann Pick disease and Alzhiemers [19] are related with the abnormality of ubiquitins. Ubiquitin conjugating enzyme, E2-EPF (M91670) was upregulated in CLN9- deficient patient cells.

In concordance with up-regulated ubiquitin pathway, proteosome subunits beta type 2, 3 and 4 were also upregulated. The proteasome is responsible for degradation of short lived and misfolded cytosolic and nuclear protein in the cells. PACE4, which encodes a subtilisin-related dipeptidase, was also unregulated.

7. DISREGULATION OF GENES IN NERVE SYSTEM DEVELOPMENT AND FUNCTION

7.1 Neuronal Development

Cytosolic platelet-activating factor acetylhydrolase gamma-subunit (PAF-AHG) is the catalytic subunit of platelet-activating factor acetylhydrolase which plays a role in phosphocholine metabolism. By comparison of the expression pattern of PAH-AHG between adult and fetal tissues, previous studies suggest that the enzyme plays an important role during the development of the brain [20]. PAF-AHG (D63391) was disregulated by +4 fold in CLN9.

Mutations in the KAL1 gene cause the X-linked Kallmann syndrome. It was suggested that KAL1 functions in late neuronal differentiation and/or neuronal trophicity [21], neural cell adhesion, and axonal migration [22]. KAL1 was upregulated 20.3 fold in CLN9.

7.2 Neuronal Cell Proliferation

Neuroendocrine DLG (NEDLG, U49089) was upregulated 10.8 fold in CLN9. Mutations of the Drosophila homologue ‘discs large’ (DLG) tumor suppressor locus lead to imaginal disc neoplasia and a prolonged larval period followed by death. Drosophila DLG and related proteins form a subfamily of the membrane-associated guanylate kinase (MAGUK) protein family and are important components of specialized cell junctions. Northern blot analysis revealed that human NEDLG is highly expressed in neuronal and endocrine tissues. Immunolocalization studies indicate that the protein was expressed mainly in nonproliferating cells, such as neurons. It was suggested that NEDLG might negatively regulate cell proliferation through its interaction with the APC protein [23].

RTVP1 (X91911) is a glioma pathogenesis-related protein. Multiple mRNAs encoded by RTVP-1 were highly expressed in a panel of cell lines from nervous system tumors arising from glia, although expression was low or absent in nonglial-derived nervous system tumor cell lines [24]. RTVP1 was downregulated 11.6 fold in CLN9.

7.3 Neuronal Electrical Excitability

KCNQ2 (D82346) is a voltage-gated potassium channel that contributes to M-channel regulation of neuron electrical excitability [25]. Mutation of KCNQ2 causes benign

neonatal epilepsy [26]. KCNQ2 was upregulated 29.1 fold in CLN9.

CHRNA5 (M83712 and U62423) is a neuronal acetylcholine receptor. It is mapped to 15q24, which is one of the loci of autosomal dominant nocturnal frontal-lobe epilepsy [27]. KCNQ2 was upregulated 14.1 fold in CLN9.

8. FUTURE WORKS

The initial success of mining knowledge from a large amount of microarray data demonstrates the necessity, the feasibility, and the potential of applying KDD in genomics. However, there are many research issues that need extensive study to make KDD in biological databases a reality, as illustrated below:

- Data from high-throughput genomic techniques are error-prone. The handling of noisy and uncertain information can be a significant problem in real-world applications. Modeling genomic information using probability, fuzzy logic, and rough set approaches are important research issues.
- The real-world databases and knowledge bases tend to be dynamic, incomplete, redundant, and noisy. Mining from multiple sources that are distributed over Internet and furnished with potentially conflicting information are open problems for future research in bioinformatics.
- There are a lot of data-mining tools from researches in statistics, pattern recognition, and machine learning. Selecting the appropriate tool to answer meaningful questions in biology is still in its infancy. This task requires more collaboration between biologists and computer scientists.

In summary, we believe that KDD is an attractive and challenging research field that will be beneficial in genomic discovery in the years to come.

9. ACKNOWLEDGMENTS

Thanks to Dr. Holly Dressman at the Duke Genome Technology Center for the Affymetrix chip hybridization, Kimberley Johnson at the Cancer Center Information Systems for helpful discussions, and Patrick McConnell for editing the manuscript.

10. REFERENCES

1. Dawson, G. and S. Cho, *Batten's disease: clues to neuronal protein catabolism in lysosomes.* J Neurosci Res, 2000. **60**(2): p. 133-40.
2. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
3. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [see comments].* Nature, 2000. **403**(6769): p. 503-11.
4. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proc Natl Acad Sci U S A, 1999. **96**(12): p. 6745-50.
5. Fayyad, U.M., *Advances in knowledge discovery and data mining*. 1996, Menlo Park, Calif.: AAAI Press : MIT Press. xiv, 611.
6. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
7. Ly, D.H., et al., *Mitotic misregulation and human aging.* Science, 2000. **287**(5462): p. 2486-92.
8. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
9. Hauser, M. and S.M. Lin, *Gene Expression Profile Analysis*, in *Approaches to Gene Mapping in Complex Human Diseases*. 2001, Wiley-Liss: New York.
10. Bassett, D.E., Jr., M.B. Eisen, and M.S. Boguski, *Gene expression informatics--it's all in your mine.* Nat Genet, 1999. **21**(1 Suppl): p. 51-5.
11. Kohonen, T., *Self-organizing maps*. Springer series in information sciences ; 30. 1995, Berlin ; New York: Springer. xv, 362.
12. Gelbard, H.A., R.M. Boustany, and N.F. Schor, *Apoptosis in development and disease of the nervous system: II. Apoptosis in childhood neurologic disease.* Pediatr Neurol, 1997. **16**(2): p. 93-7.
13. Kamada, S., et al., *Involvement of caspase-4(-like) protease in Fas-mediated apoptotic pathway.* Oncogene, 1997. **15**(3): p. 285-90.
14. Brinkmann, U., et al., *Role of CAS, a human homologue to the yeast chromosome segregation gene CSE1, in toxin and tumor necrosis factor mediated apoptosis.* Biochemistry, 1996. **35**(21): p. 6891-9.
15. Tyynela, J., et al., *A mutation in the ovine cathepsin D gene causes a congenital lysosomal storage disease with profound neurodegeneration.* Embo J, 2000. **19**(12): p. 2786-92.
16. Mikoshiba, K., *The InsP3 receptor and intracellular Ca2+ signaling.* Curr Opin Neurobiol, 1997. **7**(3): p. 339-45.
17. Cuervo, A.M. and J.F. Dice, *A receptor for the selective uptake and degradation of proteins by lysosomes.* Science, 1996. **273**(5274): p. 501-3.
18. Lee, D.H. and A.L. Goldberg, *Proteasome inhibitors: valuable new tools for cell biologists.* Trends Cell Biol, 1998. **8**(10): p. 397-403.
19. Lowe, J., et al., *Ubiquitin is a common factor in intermediate filament inclusion bodies of diverse type in man, including those of Parkinson's disease, Pick's disease, and Alzheimer's disease, as well as Rosenthal fibres in cerebellar astrocytomas, cytoplasmic bodies in muscle, and*

mallory bodies in alcoholic liver disease. J Pathol, 1988. **155**(1): p. 9-15.

20. Adachi, H., et al., *cDNA cloning of human cytosolic platelet-activating factor acetylhydrolase gamma-subunit and its mRNA expression in human tissues.* Biochem Biophys Res Commun, 1995. **214**(1): p. 180-7.
21. Hardelin, J.P. and C. Petit, *A molecular approach to the pathophysiology of the X chromosome-linked Kallmann's syndrome.* Baillieres Clin Endocrinol Metab, 1995. **9**(3): p. 489-507.
22. Legouis, R., et al., *The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules.* Cell, 1991. **67**(2): p. 423-35.
23. Makino, K., et al., *Cloning and characterization of NE-dlg: a novel human homolog of the Drosophila discs large (dlg) tumor suppressor protein interacts with the APC protein.* Oncogene, 1997. **14**(20): p. 2425-33.
24. Rich, T., et al., *RTVP-1, a novel human gene with sequence similarity to genes of diverse species, is expressed in tumor cell lines of glial but not neuronal origin.* Gene, 1996. **180**(1-2): p. 125-30.
25. Singh, N.A., et al., *A novel potassium channel gene, KCNQ2, is mutated in an inherited epilepsy of newborns.* Nat Genet, 1998. **18**(1): p. 25-9.
26. Biervert, C., et al., *A potassium channel mutation in neonatal human epilepsy.* Science, 1998. **279**(5349): p. 403-6.
27. Phillips, H.A., et al., *Autosomal dominant nocturnal frontal-lobe epilepsy: genetic heterogeneity and evidence for a second locus at 15q24.* Am J Hum Genet, 1998. **63**(4): p. 1108-16.