Candidate Gene Analysis of the Price Foundation Anorexia Nervosa Affected Relative Pair Dataset

A.W. Bergen1,2‡, M.Yeager1,2, R. Welch1,2, J.K. Ganjei1, A. Deep-Soboslay1,3, K. Haque1,2, M.B.M. van den Bree1,4, D. Goldman1,5, W.H. Berrettini6, W.H. Kaye1,7* and the Price Foundation Collaborative Group (www.anbn.org)

1Biognosis U.S., Inc. (Dissolved), Gaithersburg, MD 20877, USA, 2Currently at the Core Genotyping Facility, Advanced Technology Center, National Cancer Institute, Gaithersburg, MD 20877, USA, 3Currently at the Clinical Brain Disorders Branch, NIMH, NIH, Bethesda, MD 20892, USA, 4Currently at the Division of Psychological Medicine, University of Wales College of Medicine, Cardiff, United Kingdom, 5Laboratory of Neurogenetics, NIAAA, NIH, Rockville, MD 20852, USA, 6Center for Neurobiology and Behavior, University of Pennsylvania, Philadelphia, PA 19104, USA, 7Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15213, USA

Abstract: The eating disorders are severe psychiatric illnesses with significant morbidity and mortality that exhibit statistically significant familial risk and heritability, providing support for a molecular genetic approach toward defining etiological factors. An emerging candidate gene literature has concentrated on serotonergic and dopaminergic candidates. With the financial support of the Price Foundation, a group of investigators initiated an international multi-center collaboration (Price Foundation Collaborative Group) in 1995 to study the genetics of anorexia and bulimia nervosa by collecting and analyzing phenotypes and genotypes of individuals and their relatives affected with eating disorders. The first sample of families collected by this collaborative group, known as the Price Foundation Anorexia Nervosa Affected Relative Pair (AN-ARP) dataset, was ascertained on an proband affected with Anorexia Nervosa (AN), with relative pairs affected with the eating disorders AN, Bulimia Nervosa or Eating Disorders Not Otherwise Specified [1].

Biognosis U.S., Inc. was founded to identify and characterize candidate susceptibility genes for anorexia and bulimia nervosa phenotypes in the Price Foundation eating disorder datasets. During 2000-2001, Biognosis U.S., Inc. developed and implemented a research program with a focus on the analysis of candidate gene polymorphisms [3], neuroendocrine regulation of appetite [4], and by a positional hypothesis from a linkage analysis of the AN-ARP dataset [5]. This report reviews the anorexia nervosa candidate gene literature through 2001, the candidate gene research program implemented at Biognosis U.S., Inc. and selected candidate gene findings in the AN-ARP dataset derived from that research program.

Key Words: Anorexia Nervosa, Candidate Gene, Indolamine, Catecholamine, Neuropeptide, Genotypes, Alleles, Haplotypes, Transmission Disequilibrium Test.

INTRODUCTION

Anorexia Nervosa Phenotype

Anorexia nervosa ("AN") affects ~ 0.5% of women [6] and carries a significant mortality risk [7]. Although the majority of patients (70-80%) do recover, this may take considerable time (5 – 6 years) [8]. The diagnostic criteria for AN include: refusal to maintain body weight at or above a minimally normal weight, fear of gaining weight even though underweight, and a disturbance in the patient's perception of body weight or shape and its effect on self-evaluation, and amenorrhea [9]. Comorbidity of AN and other psychiatric disorders (e.g., depression, OCD, anxiety disorders, bipolar disorder) and extremes of personality traits have been described [10-14]. The reported familial relative risk is approximately 11 for AN [15], suggesting genetic factors may play a role. This is confirmed by twin studies, which have reported heritability rates ranging from 50 and 80% [16-20]. These findings provide the basis for a molecular genetic approach to the study of eating disorder susceptibility factors.

Published molecular genetic association studies on AN through 2001 are reviewed below and in [21, 22]. Candidate gene association analyses to AN phenotype at sixteen genes have been published in the English language literature as of April 2001 (Table I and Table II). These analyses compare genotype and allele counts from one or more DNA polymorphisms at candidate genes in individuals affected
with AN (cases) versus individuals unaffected with AN (controls), i.e., case:control contingency association analysis. Case status was either DSM-IIIR or DSM-IV AN. Typically, candidate gene case:control studies in eating disorder samples have been based on a single polymorphism at most of these genes, although two serotonin loci have received attention from several groups and several groups have screened a candidate gene for novel polymorphisms (Table I). The mean AN sample size in these candidate gene studies is approximately 90 individuals, while the mean control sample is about 180 individuals. Control samples in this candidate gene literature are very diverse, consisting of both psychiatrically screened and unscreened samples, normal weight, underweight and obese samples, and samples of both sexes. Several groups of investigators have samples that include parents or unaffected siblings. Family samples permit both association analysis and linkage analysis and reduce the probability of type I error (false positives) at the cost of increased type II error (false negatives). Ideally, any positive association findings in case:control samples must be evaluated in additional case:control samples and in family samples to increase confidence in a proposed relationship between candidate gene sequence variation and susceptibility to an eating disorder.

Serotonergic Candidate Gene Studies Through 2001

Six serotonergic candidate gene loci have been the subject of publications in the AN candidate gene literature, i.e., the genes encoding the serotonin receptors 1B, 2A, 2C and 7, the serotonin transporter, and tryptophan hydroxylase, the rate limiting enzyme for serotonin synthesis in brain (Table I). While substitution polymorphisms at the loci for serotonin receptors 1B (Phe124Cys) and 7 (Pro279-Leu) have been previously associated with alcoholism [23] and antisocial personality disorder [24], the result of association studies between these candidate polymorphisms and AN was non-significant [25]. A sequence polymorphism flanking the HTR2A locus (-1438G>A) has been reported to be significantly associated with both AN and OCD [26, 27], and with OCD in females only [28]. However, some studies of this HTR2A polymorphism have not exhibited statistically significant association to AN, e.g., [29-31], and see comments to [26]. A functional serotonin 2C receptor locus substitution polymorphism (Cys23Ser) [32] has been previously associated with hyperphagia and auditory hallucinations in Alzheimer’s disease [33]; but this substitution was not found to be significantly associated to AN [31]. The serotonin transporter gene locus (SLC6A4) has a functional promoter polymorphism associated with transcription efficiency [34] that has been associated with anxiety-related traits [35]. Several studies with this polymorphism do show statistically significant association to AN [36, 37], while some studies fail to show association [38, 39]. A screen of the tryptophan hydroxylase locus (TPH) did not identify any polymorphisms associated to AN [40].

Catecholamine Candidate Gene Studies Through 2001

Polymorphisms at three dopaminergic genes have been investigated for association to AN in the literature, i.e.,

| Table I. Serotonergic Candidate Gene Association Studies in AN Through 2001 |
|-----------------|---------|-------|--------|--------|--------|-----|--------|
| Gene Locus     | OMIM#   | Chr   | Poly.  | Ca. N  | Cn. N  | Result | Citation       |
| HTR1B          | 182131  | 6q13  | Phe124Cys | 84     | 393    | NS    | Hinney, '99   |
| HTR2A          | 182135  | 13q14-q21 | -1438G/A | 78     | 170    | NS    | Ziegler, '99  |
|                |         |       |        | 81     | 226    | S     | Collier, '97  |
|                |         |       |        | 100    | 295    | NS    | Hinney, '97   |
|                |         |       |        | 152    | 150    | S     | Campbell, '98 |
|                |         |       |        | 77     | 107    | S     | Sorbi, '98    |
|                |         |       |        | 68     | 69     | S     | Enoch, '98    |
|                |         |       |        | 45     | 45     | NS    | Karwautz, '01 |
|                |         |       |        | 109    | 107    | S     | Nacmias, '01  |
| HTR2C          | 312861  | Xq24  | Cys23Ser | 45     | 45     | NS    | Karwautz, '01 |
| HTR7           | 182137  | 10q21-q24 | Pro279Leu | 84     | 393    | NS    | Hinney, '99   |
| SLC6A4         | 182138  | 17q11.1-q12 | PromoterVNTR | 138    | 90     | NS    | Sundaramurth, '00 |
|                |         |       |        | 56     | 120    | S     | Di Bella, '00 |
|                |         |       |        | 67     | 148    | NS    | Fumeron, '01  |
|                |         |       |        | 67     | 358    | S     | Fumeron, '01  |
|                |         |       |        | 96     | 385    | NS    | Hinney, '97   |
| TPH            | 191060  | 11p15.3-11.4 | C1095T | 128    | 142    | NS    | Han, '99      |
DRD3, DRD4 and COMT (Table II). Previously, these genes have been associated to schizophrenia (DRD3 receptor) [41], novelty-seeking and substance abuse (D4 receptor, inconsistent findings) [42], and OCD (COMT) [43]. The published reports in the eating disorder literature do not identify an association to AN at polymorphisms at these genes in case:control studies [44, 45, 29]. Two family-based association studies using a functional substitution polymorphism at COMT (Val158Met) present evidence for genes in case:control studies [44, 45, 29]. Two family-based association to AN [48].

Neuroendocrine Candidate Gene Studies Through 2001

A series of appetite regulation and energy metabolism candidate genes have been evaluated for association with AN (Table II). These candidate genes have been selected using the abundant evidence that these genes are involved in the control of appetite and feeding behavior [4]. A case:control study has provided evidence that a two locus polymorphism that includes a substitution polymorphism at the agouti related protein locus is associated with AN [49]. DNA polymorphisms at the following neuroendocrine loci were all studied for associations with AN: leptin, the adipocyte orexigenic hormone; melanocortin receptor-4, a receptor regulating feeding with agonist NPY and antagonist AGRP; the NPY Y5 receptor; and proopiomelanocortin, the locus coding for corticotropins, lipotropins, melanotropins and beta-endorphin. For all of the polymorphisms tested, no statistically significant association was found with AN [50-53]. The uncoupling proteins UCP1, UCP2, UCP3 have previously been associated with energy expenditure related traits [54], and a microsatellite marker linked to the UCP2/3 gene cluster on chr11q13 has shown significant association to AN [55], however the significance of this finding should be evaluated by association analysis of variation at the genes themselves. The estrogen receptor beta gene, a polymorphism at which has been associated with ovulatory dysfunction [56], has also been shown to be nominally significantly associated to AN [57]. Finally, studies at the major histocompatibility locus HLA-A have been inconclusive [58, 59].

In summary, over twenty polymorphisms at sixteen (N=16) candidate genes have been evaluated for association with the diagnoses of AN in the psychiatric genetic literature (Tables I and II). Association findings at the serotonin 2C receptor and serotonin transporter loci has been shown to reproducibly convey risk, however, the risk is modest, i.e., a RR of ~1.4 fold and ~1.25 fold, respectively. Why are the results of most of these candidate gene association studies to AN phenotype negative or equivocal despite the use of neurobiological hypotheses to nominate candidates? The fundamental reasons are 1) the very low prior probability of any particular candidate gene being associated with any complex disorder [60], due to the large number of genes (>30,000) and sequence polymorphisms (>3,000,000) known to exist in the genome [61], 2) the large number of individuals required to detect the modest genotypic relative risks expected in a complex disorder such as AN in association and linkage designs [62], and the possibility that the existing psychiatric diagnostic criteria of AN may not be the best phenotype to identify genetic susceptibility factors known to exist in the genome [61], 2) the large number of genes in AN phenotype negative or equivocal despite the use of neurobiological hypotheses to nominate candidates? The fundamental reasons are 1) the very low prior probability of any particular candidate gene being associated with any complex disorder [60], due to the large number of genes (>30,000) and sequence polymorphisms (>3,000,000) known to exist in the genome [61], 2) the large number of individuals required to detect the modest genotypic relative risks expected in a complex disorder such as AN in association and linkage designs [62], and the possibility that the existing psychiatric diagnostic criteria of AN may not be the best phenotype to identify genetic susceptibility factors in the eating disorders [63]. In addition, samples are often ascertained in widely variable and unsystematic ways, and may differ with respect to a variety of factors influencing outcome. These could include, for example, the presence of comorbid traits and other modifying factors that influence disease severity.

<table>
<thead>
<tr>
<th>Locus</th>
<th>OMIM#</th>
<th>Chr</th>
<th>Poly.</th>
<th>Ca. N</th>
<th>Cn. N</th>
<th>Result</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRD3</td>
<td>125451</td>
<td>3q13.3</td>
<td>Ser9Gly</td>
<td>39</td>
<td>42</td>
<td>NS</td>
<td>Bruins-Slot, '98</td>
</tr>
<tr>
<td>DRD4</td>
<td>126452</td>
<td>11p15.5</td>
<td>Ex1del,Ex3VNTR</td>
<td>109</td>
<td>327</td>
<td>NS</td>
<td>Hinney, '99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ex1del,Ex3VNTR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ex3 VNTR</td>
<td>45</td>
<td>45</td>
<td>NS</td>
<td>Karwautz, '01</td>
</tr>
<tr>
<td>COMT</td>
<td>116790</td>
<td>22q11.2</td>
<td>Val158Met</td>
<td>45</td>
<td>45</td>
<td>NS</td>
<td>Karwautz, '01</td>
</tr>
<tr>
<td>AGRP</td>
<td>602311</td>
<td>16q22</td>
<td>760G/A, 526G/A</td>
<td>145</td>
<td>244</td>
<td>S</td>
<td>Vink, '01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>830C/T</td>
<td>145</td>
<td>244</td>
<td>NS</td>
<td>Vink, '01</td>
</tr>
<tr>
<td>LEP</td>
<td>164160</td>
<td>7q31.3</td>
<td>-1387G/A</td>
<td>115</td>
<td>315</td>
<td>NS</td>
<td>Hinney, '98</td>
</tr>
<tr>
<td>MCH4</td>
<td>155541</td>
<td>18q22</td>
<td>Val1031e</td>
<td>51</td>
<td>52</td>
<td>NS</td>
<td>Hinney, '99</td>
</tr>
<tr>
<td>UCP</td>
<td>601693</td>
<td>11q13</td>
<td>D11S911</td>
<td>170</td>
<td>150</td>
<td>S</td>
<td>Campbell, '99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D11S916</td>
<td>170</td>
<td>150</td>
<td>NS</td>
<td>Campbell, '99</td>
</tr>
<tr>
<td>ESR1</td>
<td>601663</td>
<td>14q</td>
<td>1082G/A</td>
<td>50</td>
<td>96</td>
<td>S</td>
<td>Rosenkranz, '98</td>
</tr>
<tr>
<td>HLA-A</td>
<td>142800</td>
<td>6q21.3</td>
<td>Bw16 et al.</td>
<td>37</td>
<td>246</td>
<td>S</td>
<td>Biederman, '84</td>
</tr>
</tbody>
</table>
AN-ARP DATASET COLLECTION AND ANALYSIS

In 1995, a multi-center collaborative study on the genetics of AN and bulimia nervosa (BN) was organized by Walter H. Kaye as Principal Investigator, with the financial support of the Price Foundation of Geneva, Switzerland. Data collection has been performed at multiple clinical sites, with a group of clinically administered and self-report questionnaires, using the Structured Interview for Anorexia and Bulimia to make a DSM-IV diagnosis of AN [2]. A sample of blood was collected from each individual for genetic studies. The sample of eating disorder affected relative pairs ascertained on a AN proband is known as the Price Foundation Anorexia Nervosa Affected Relative Pair dataset (AN-ARP) [2]. The affected relatives fulfill criteria for DSM-IV anorexia nervosa (AN), bulimia nervosa (BN) or eating disorder not otherwise specified (ED NOS). A second sample of eating disorder affected relative pairs was ascertained on a BN proband, collected, and is known as the Price Foundation Bulimia Nervosa Affected Relative Pair dataset (BN-ARP) [64]. Blood samples from parents have been collected to provide additional genetic information for gene mapping studies. The first ARP dataset collection was completed in 1998 and the second was completed in 2000.

A Weber version 9 (Marshfield Medical Foundation Center for Medical Genetics) whole genome linkage scan has been applied to DNA samples from both datasets [2, 64]. Several studies of the behavioral, clinical and temperament phenotypes in the AN-ARP dataset, as well as three linkage analyses of the two genome scans, are extant [5,64-69].

Biognosis U.S., Inc.

Biognosis U.S., Inc. ("Biognosis") was founded in 1999 to identify and characterize candidate genes in the Price Foundation eating disorder datasets. The founding group of scientific advisors included scientists from the Price Foundation Collaborative Group and scientists with expertise in biological psychiatry, psychiatric genetics, candidate gene analysis and human molecular genetic analysis. The scientific mission of the company was to identify genes increasing susceptibility to eating disorders. Biognosis created a molecular genetic analysis capability based on Applied Biosystems (Foster City, CA) sequencing and genotyping technology platforms and the implementation of a LabVantage (Bridgewater, NJ) Laboratory Information Management System (LIMS). Biognosis outfitted a laboratory for sequencing and genotyping and implemented a LIMS in early 2000. Biognosis U.S., Inc. was wound up in December of 2001. The molecular genetic analysis capability developed in that period was used to investigate candidate genes in the Price Foundation eating disorder datasets and is described below.

Development of a Molecular Genetic Analysis Capability

Biognosis implemented the following molecular genetic methods in its Gaithersburg, MD laboratory facility to investigate the Price Foundation AN-ARP and BN-ARP datasets. Blood and DNA specimens were received, and tracking of these samples was initiated within the LIMS. DNA from blood samples from individuals whose DNA was not already available was extracted using a standard high-salt extraction procedure. Newly extracted DNA and other DNA samples received into the laboratory were quantified by measurement of UV absorbance at 260 nm. DNA was diluted, aliquoted, and normalized to 10 ng/μL based on this quantification information and assessment of a test PCR. Aliquots were first plated onto master 96-well plates and subsequently onto working plates (50ng/well), then lyophilized for later use in genotyping assays. Blood samples received as duplicates were stored appropriately and extracted at a later date when it was determined that DNA was needed.

The selection of candidate genes to investigate in the Price Foundation AN-ARP dataset was based on neurobiological and positional hypotheses [2-5]. Characterization of these candidate genes and optimization of genotyping assays for selected genetic variation within these genes was based on bioinformatic characterization of sequence and polymorphism information from public databases. Genomic and mRNA sequence and polymorphism information was downloaded from public databases and obtained from the biomedical literature and assembled to form a genomic sequence contig with annotation at structural and variable positions. Information concerning variation at these candidate loci was incorporated into the LIMS and prioritized for validation and genotype assay development. In cases where genomic variation at a candidate gene was not available in public databases or the literature, or the information concerning the variation was incomplete, a portion of the gene or the entire gene, was sequenced in order to identify or validate variable sites using PCR-based DNA resequencing. PCR primers were designed to produce 300-600 bp products that were resequenced in 32 unrelated CEPH individuals using Applied Biosystems (Foster City, CA) sequencing reagents, automated sequencers, and sequencing analysis software. This additional information was incorporated into the LIMS and used to select variable sites (mostly single nucleotide polymorphisms or SNPs) to genotype in the Price Foundation AN-ARP and control datasets. Decisions on which candidate gene SNPs to genotype were based on the criteria of a priori findings in the literature, allele frequency, possible substitution effects, and location of the polymorphism within a candidate gene, derived from the bioinformatic analysis of public and Biognosis generated sequence information.

Biognosis utilized four different genotyping platforms: TaqMan™ 5’ exonuclease, fluorogenic primer extension genotyping (SNap-Shot™), short tandem repeat genotyping by fluorescent PCR size determination using gel electrophoresis, and PCR-RFLP analysis. The majority of genotyping assays were designed as TaqMan 5’ exonuclease assays. For the 5’ exonuclease assays, dual-labeled probes and primers were chosen using ProbeITY (Celadon Laboratories, College Park, MD) and synthesized by Applied Biosystems. Assays were optimized using standard conditions [70], by using control DNAs with known genotypes for a single nucleotide polymorphism (SNP), and
by modifying assay conditions until an assay produced those known genotypes accurately, as measured by the criteria of completion and concordance in a duplicate sample of CEPH DNAs. Once an assay was optimized, it was applied to case, family and control DNA samples. Genotype determination was conducted manually using Applied Biosystems software on the respective Applied Biosystems instrument (Sequence Detector 7700 or Automated Sequencer 377XL). For most assays, a verification plate consisting of a 17% duplication of the appropriate data set and control group samples was genotyped at the same time in order to assess genotype concordance rate. Genotyping error rate was assessed by the use of no template (no genomic DNA) and known genotype controls, and the criteria of completion rate, concordance rate, and Hardy Weinberg Equilibrium (HWE) contingency analysis.

**Development of an Informatics Infrastructure**

From October through December of 1999, Biognosis investigated existing LIMS vendors and software applications for integration into the laboratory. Biognosis executed a license with a LabVantage Solutions for its Sapphire LIMS software in December of 1999 and implemented a LIMS throughout 2000. The Biognosis LIMS was based on an Oracle database, extensive use of VBScripts and built-in client side workflows. This LIMS was designed and constructed to assist the laboratory in three major areas: sample and reagent tracking, clinical, bioinformatic and genotype databasing, and data analysis and reporting.

In the sample tracking portion of the Biognosis LIMS, users tracked samples and information about those samples throughout all stages of laboratory use. This included sample tracking from multiple sites into multiple storage sites, and aided in the selection of samples to be used for the sample extraction workflow, and queries based on the current stage of the samples. The LIMS also tracked individual samples as they were transferred to aliquots on laboratory plates. Plate creation and tracking in any laboratory workflow followed the laboratory protocols that created and used various types of plates, including master plates, working plates, and working dilution plates. Groups of plates, for example, could be defined as an assay and selected for further analysis in a particular workflow. These workflows included extraction, plate creation, plate replication, PCR and genotyping assay set-up workflows.

In the databasing portion of the Biognosis LIMS, clinical, bioinformatic and genotypic data attributes were tracked and recorded. The LIMS managed clinical data, i.e., phenotype and pedigree data of the AN and BN ARP datasets. The LIMS performed candidate gene databasing including the locus and polymorphism data storage, primer and probe tracking and data storage, as well as information and analysis related to genetic markers. This enabled workflows to be incorporated allowing for automated data entry for TaqMan™ exports. For TaqMan™, users initiated a menu driven process to select the desired file, then the workflow imported the data, associated it with individual samples, and attached the raw data file to the plate.

The Biognosis LIMS also assisted in data analysis of genotypes and association between genotypes and phenotypes. The LIMS had routines that automatically calculated dropout rates as a count of failed genotypes, flagged plates based on % completion per plate, measured discordance between duplicated samples, tested HWE and performed contingency table (association) analysis on alleles and genotypes, after the user selected a series of plates to be analyzed. The LIMS also enabled reporting (exporting) of data in various formats for a number of external analytical programs, e.g., Linkage, TDT, EM, Arlequin and Excel formats as well as producing statistical reports. These statistical reports ranged from simple reports, such as reporting the numbers of plates in the LIMS used for each of the various genotyping assays, to contingency analyses of genotypes and phenotypes. Candidate gene polymorphisms that were identified as nominally associated by contingency analysis in this manner were further investigated using statistical analysis as described below using data exported from the LIMS.

**Statistical Analysis of Candidate Gene Genotypes**

Multi-locus genotypes were exported from the LIMS and assembled in a Visual Basic utility and resulting multilocus genotype counts were used to estimate intragenic and intergenic pairwise linkage disequilibrium using likelihood ratio tests with empirical significance testing using Arlequin [71]. The significance of differences in genotype and allele frequencies at the polymorphic loci between the AN and control samples was evaluated using contingency analysis in SAS. In case of expected cell frequency ≤5, the p-value was based on Fisher's exact test. The empirical significance of haplotype frequency differences between AN and control samples was evaluated using the nonparametric heterogeneity statistic (T5) in the program EH+ using 10,000 permutations [72]. In order to maximize the power to detect transmission disequilibrium at individual SNPs, FBAT was used for transmission disequilibrium analysis (TDT) [73]. Power and the required sample size were evaluated using chi-square distribution properties and, for the TDT, the pedigree structures. A p value of ≤0.05 is described in all analyses as indicating a statistically significant result, while p values >0.05 and <0.10 are described in all analyses to indicate a trend towards statistical significance.

**Control DNA Samples**

DNA samples collected from a group of European-American females (EAF) were used as control DNA samples for genotype and allele frequency comparisons in the case:control design. The EAF sample consists of European-American females (N=125), recruited through advertisements, screened to exclude obese individuals (>120% ideal weight) and to exclude lifetime criteria for Axis I disorders as assessed by clinical administration of the Structured Clinical Interview for DSM-III-R (SCID) [74]. Two publicly available sets of DNA samples were utilized for SNP discovery and validation and for quality control samples of known genotype. These two sets of samples were the Centre Etude Polymorphism Humaine (CEPH) DNA
samples and the Coriell Caucasian Human Diversity Panel (COR), obtained from the Coriell Cell Repositories. CEPH DNA was used primarily for resequencing and as control DNAs for genotype assay development and for genotyping. The CEPH DNA samples are DNA samples extracted from cell lines of a set of American, French and Venezuelan pedigrees collected for the purpose of recombination mapping by a collaboration organized by Jean Dausset of CEPH. The CEPH samples used at Biognosis consisted of unrelated individuals. COR DNA samples were used as an additional DNA sample for various quality control purposes. The COR sample consists of 100 DNA samples (N=49 females, N=51 males) from cell lines maintained at the Coriell Cell Repository from self-identified Caucasians.

Candidate Gene Analysis in the AN-ARP Dataset

Biognosis' molecular genetic analysis strategy was to investigate candidate genes whose products are involved in monoamine and neuropeptide neurotransmission and in the central nervous system control of appetite regulation. These genes are nominated as candidates in AN through studies investigating CSF levels of monoamine and neuropeptides, in models of AN which emphasize behavioral reward reinforcement and in animal models of anxiety, appetite regulation and eating behavior. The following general hypotheses regarding the biological basis of AN were applied to nominate candidate genes for investigation. A leading hypothesis is that serotonergic dysfunction, revealed in studies of CSF serotonergic metabolite differences between eating disorder patients before and after recovery, and also between recovered eating disorder patients and controls, increases vulnerability to eating disorders [75]. Another leading hypothesis is that the stress of self-starvation and exercise, by increasing the release of corticotropin releasing factor from the hypothalamus, activates dopaminergic, adrenergic and opioidergic pathways [76]. Indeed, two studies of concentrations of opioid peptides have shown differences in individuals with AN versus controls [77, 78]. Those individuals vulnerable to reward reinforcement might, according to this hypothesis, become addicted to the reward produced by the activation of the Hypothalamic Pituitary Adrenal axis [79]. Another class of hypotheses concerns central nervous system (CNS) control of food intake, which resides primarily within the hypothalamus and has been categorized into gene products that increase food intake (orexigenic) and those that decrease food intake (anorexigenic) [4]. Many neuropeptides and receptors are involved in this feedback loop, and several reward reinforcement candidates are also involved in the appetite regulation. Finally, a positional hypothesis derived from a linkage scan of the AN-ARP dataset helped to prioritize several candidates for molecular genetic investigation. Specifically, genome-wide linkage analysis of the entire AN-ARP dataset provided only modest evidence for linkage when all DSM-IV eating disorder diagnoses were included in the affected sample [5]. However, linkage analysis in a subset of 37 families with at least one affected relative pair concordantly affected with a DSM-IV diagnosis of restricting AN (a subtype of AN), provided suggestive evidence for linkage with a maximum multipoint NPL score of 3.45 found on chromosome 1p33-36 at marker D1S3721, located at ~72.6cm on the Marshfield linkage map [5].

Analysis of OPRD1 and HTR1D

Results from two examples of candidate genes with statistically significant association to AN phenotype in the Price Foundation AN-ARP dataset, i.e., the delta opioid and serotonin 1D receptor, OPRD1 and HTR1D [80], are reviewed below. OPRD1 and HTR1D are candidate genes for AN nominated by multiple hypotheses: positional (both genes are located on chromosome 1p33-36, a region linked to AN [5]; the anxiety and appetite modulation functions of the OPRD1 gene product; and the autoreceptor function of the HTR1D gene product. These two candidate loci were

<table>
<thead>
<tr>
<th>SNP1</th>
<th>Allele 1</th>
<th>Allele 2</th>
<th>SNP ID2</th>
<th>Coding region</th>
<th>Source</th>
<th>AI3</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTR1D(-1123T&gt;C)</td>
<td>T</td>
<td>C</td>
<td>SNP000083015</td>
<td>No, 5’ of coding</td>
<td>Reference 80</td>
<td>31.3</td>
</tr>
<tr>
<td>HTR1D(-628T&gt;C)</td>
<td>T</td>
<td>C</td>
<td>SNP000083091</td>
<td>No, 5’ of coding</td>
<td>Reference 80</td>
<td>14.1</td>
</tr>
<tr>
<td>HTR1D(1080C&gt;T)</td>
<td>C</td>
<td>T</td>
<td>SNP00006432</td>
<td>Yes, silent</td>
<td>Reference 83</td>
<td>9.4</td>
</tr>
<tr>
<td>HTR1D(2190A&gt;G)</td>
<td>A</td>
<td>G</td>
<td>SNP000082070</td>
<td>No, 3’ of coding</td>
<td>Reference 80</td>
<td>65.6</td>
</tr>
<tr>
<td>OPRD1(807T&gt;G)</td>
<td>T</td>
<td>G</td>
<td>SNP000063484</td>
<td>Yes, F27C</td>
<td>Reference 82</td>
<td></td>
</tr>
<tr>
<td>OPRD1(8214T&gt;C)</td>
<td>T</td>
<td>C</td>
<td>SNP001026473</td>
<td>No, IVS 1</td>
<td>TSC0110129</td>
<td></td>
</tr>
<tr>
<td>OPRD1(23340G&gt;A)</td>
<td>G</td>
<td>A</td>
<td>SNP000085600</td>
<td>No, IVS 1</td>
<td>TSC0110127</td>
<td></td>
</tr>
<tr>
<td>OPRD1(47821A&gt;G)</td>
<td>A</td>
<td>G</td>
<td>SNP000066643</td>
<td>No, IVS 2</td>
<td>Reference 80</td>
<td>41.1</td>
</tr>
<tr>
<td>OPRD1(51502A&gt;T)</td>
<td>A</td>
<td>T</td>
<td>SNP000066640</td>
<td>No, 3’ of coding</td>
<td>TSC0110133</td>
<td></td>
</tr>
</tbody>
</table>

1From the ATG of the candidate gene genomic sequence.
2HGVBASE at http://hgvbase.cgb.ki.se/.
3Percent allele 1 from 32 unrelated CEPH DNAs obtained by resequencing.
identified as positional candidates through a search of unfinished genome sequencing (HTG) section of Genbank [81] using the sequence of chr1p33-36 microsatellite markers. OPRD1 and HTR1D reside at 55.7cM/33.5Mbp and 49.0cM/23.7Mbp, -23cM and 17cM distal to the linkage peak at 72.6cM [80], respectively. However, OPRD1 and HTR1D are not in significant linkage disequilibrium with each other in the AN sample [80]. Therefore, for the purposes of association analysis, these two genes can be treated as two distinct genetic targets.

Sequence Variation at OPRD1 and HTR1D SNPs

The OPRD1 locus contains three exons of 51,551 base pairs with 1,173 base pairs of cDNA sequence and the HTR1D locus contains a single exon with 1,134 base pairs of cDNA sequence. In order to validate previously reported SNPs and to identify novel SNPs, portions of both loci were resequenced in 32 CEPH individuals: 2.8Kb of the HTR1D locus (from -1165bp 5' to +1497bp) and 4.1Kb portion of the genomic sequence surrounding exon 2 of OPRD1 (from IVS1-1860 to IVS2+1996) [80]. The resequencing of 32 CEPH samples at OPRD1 and HTR1D identified one de novo OPRD1 SNP and four de novo HTR1D SNPs and validated one previously reported HTR1D SNP [82]. Nine TaqMan™ genotyping assays were developed, four at HTR1D and five at OPRD1, from de novo, previously reported and TSC in silico SNPs [Table IV] [80]. Average genotype completion and concordance rates were 95% and 98.2% for the AN proband sample and 88% and 98.6% for the EAF control sample, respectively [80].

Association of OPRD1 SNPs to AN

Three of five OPRD1 SNPs genotyped in the Price Foundation AN-ARP dataset [Table IV], OPRD1(8214T>C), OPRD1(23340A>G), and OPRD1(47821A>G), were observed to be significantly associated to AN (p<0.05) [80]. The strongest evidence was found for OPRD1(47821A>G) (p=0.01) with a OR of 1.64 (95%CI 1.11-2.44), allele 2 to allele 1 [80]. Significant differences in OPRD1 SNP haplotype frequencies between the AN proband and EAF samples was observed with OPRD1 SNP haplotype (8214T>C)/(47821A>G) (p=0.04) [80]. OPRD1 80G allele frequencies determined in the EAF sample (minor allele = 0.11) were not different to the previously reported frequency (minor allele = 0.12) in European-Americans [83] and there is no significant allele frequency difference between the AN proband and EAF control samples at this polymorphism (p=0.94). Significant allele frequency differences among six continental population groups have been previously observed at this polymorphism [83]. This suggests that the AN proband and EAF samples studied here may be ethically matched at a continental population level. This study also found a trend towards transmission disequilibrium at two OPRD1 SNPs (47821A>G and 51502A>T)(p=0.06); these two OPRD1 SNPs exhibited nominal linkage disequilibrium with each other (p=0.034) [80]. These results suggest that the OPRD1 locus may be involved in susceptibility to AN, however, this association should be investigated in additional samples using both case:control and family-based designs in order to confirm the results from the Price Foundation AN-ARP database.

Association of HTR1D SNPs to AN

One HTR1D SNP, HTR1D(1080C>T), of four genotyped in the Price Foundation AN-ARP dataset [Table IV], was observed to be significantly associated to AN (p=0.01, both genotypic and allelic) with a OR of 2.63 (95CI 1.21-5.75) [80]. Significant HTR1D SNP haplotype frequency heterogeneity between the AN proband and EAF samples was observed with the three pair-wise haplotypes containing the HTR1D(1080C>T) SNP (p<0.04), but not with the haplotype containing all four HTR1D SNP alleles [80]. While significant transmission disequilibrium was observed at three HTR1D SNPs, it was not observed at HTR1D(1080C>T) [80], probably because of the lack of power of the TDT with low minor allele frequencies [84]. These three HTR1D SNPs (-1123T>C, -628T>C and 2190A>G) were observed to be in complete linkage disequilibrium with each other, while the HTR1D(1080C>T) was not observed to be in complete linkage disequilibrium with the other three [80]. As the HTR1D(1080C>T) SNP is a silent mutation and the three remaining SNPs are 5' or 3' of the coding region, it is possible that these SNPs may be associated with differences in HTR1D gene expression or in linkage disequilibrium with other polymorphisms at the HTR1D locus. Again, this association should be investigated in additional samples using both case:control and family-based designs in order to confirm the results from the Price Foundation AN-ARP dataset.

Linkage of OPRD1 and HTR1D to AN

After incorporating the OPRD1 and HTR1D SNPs in the linkage analysis using the same sample of N=37 families with concordantly restricting AN affected sib pairs [5], the strength and significance of the linkage signal previously observed was increased substantially, by 0.46 Z and an order of magnitude in p value to 3.91 and p=0.00005, respectively [80], from the previous analysis [5]. However, OPRD1 and HTR1D are >16cM from D1S3721 and the correlation between allele sharing at the candidate SNPs and chrlp microsatellites, while substantial, does not account for the complete linkage signal [80]. Note that the addition of the candidate gene SNPs did not reduce the width of the linkage peak and introduced a secondary linkage peak with a NPL score of 3.05 at OPRD1 [80].

Power to Detect Observed Allelic Effects

The estimated total number of cases and controls necessary to have a 50% chance to detect a given allelic effect (odds ratio) at a given minor allele frequency using the case:control (50:50) design with a categorical phenotype and allele counts at a biallelic polymorphism was calculated. To detect an odds ratio of 1.6 with a minor allele frequency of
42%, that of the OPRD1(47821A>G) SNP [80], it is necessary to have ~300 individuals in a case-control study. Substantially more individuals will be required to have a 50% chance of detecting the HTR1D-1080C>T SNP effect, given an observed minor allele frequency between 10 and 15%, ~1500 individuals. Thus, it is apparent why the modest number of cases and controls in the published eating disorder candidate gene literature has generally resulted in non-significant findings, despite the careful selection of neurobiological candidate genes. This also suggests that a significantly larger sample will be required to attempt to replicate the candidate gene associations described here. Consideration of the severity of the AN diagnosis will also be important in defining a sample for follow-up of these findings, as the AN-ARP probands are chronically affected individuals from multiplex families and are likely to be more severely affected than singleton affecteds.

**Follow-up to Price Foundation AN-ARP Findings**

Several factors related to candidate gene association study design will influence association study results, whether the candidate gene association study is directed towards discovery or validation of association between candidate gene polymorphisms and AN. As candidate gene association analysis will not be limited to a single locus or polymorphism, but will be an ongoing process in which novel candidate regions, genes and polymorphisms are tested as they are nominated, statistical correction with a more significant α level will be required, just as in linkage analysis [85]. Exhaustive and systematic association studies of chromosomal regions nominated by linkage analysis may require the genotyping of hundreds of single nucleotide polymorphisms (SNPs), even in small genomic regions, in order to move from linkage analysis to the identification of a susceptibility gene, e.g., [86], increasing sample size requirements. In the case of whole-genome association scans, an α value of $5 \times 10^{-8}$ has been suggested by several authors [62, 87], but it is unlikely that 500,000 – 1,000,000 markers will ever be used in practice. Genotyping of several SNPs per gene will allow haplotypes to be estimated computationally, and thus most candidate gene haplotypes will be highly informative. An approach using SNPs primarily located at or within genes will increase the probability of identifying markers in linkage disequilibrium with functional polymorphisms [62]. Both of these factors will tend to reduce required sample sizes. With a relative risk estimate of 11 for AN [15], one of a number of oligogenes may have a genotypic risk ratio (GRR) of 1.5 to 2.0. If preliminary statistical analysis suggests that more than one allele is associated with increased risk, e.g., as for hemochromatosis [88], these alleles could be lumped together to create an at-risk group of alleles. However, without a priori knowledge of allelic effects, multiple alleles at a locus may increase the required sample size (depending upon frequencies and GRRs of the alleles). The above considerations suggest that in order to pursue large-scale association studies with AN phenotypes using either the case:control or the trio design, a AN sample size several times greater than the Price Foundation AN-ARP dataset is required.

One approach towards estimating required sample sizes suggests that a sample of ~500-800 trios with a proband affected by AN and one or two parents would provide a sample powerful enough to pursue the range of association strategies outlined above [62, 84]. This sample size would be sufficient to detect GRRs ~2.0 in both discovery and validation oriented candidate gene association studies, including the systematic analysis of genomic regions.
nominated by linkage analysis [5, 64, 69]. However, if genetic susceptibility to eating disorders reflects the operation of hundred of genes of small effect, even large-scale studies will be underpowered. Fortunately, the Price Foundation’s scientific advisors anticipated the necessity for the collection of a large association dataset in 1999 and made a recommendation for such a collection. This recommendation resulted in the collection of a third Price Foundation Collaborative Group clinical sample known as the Price Foundation AN Trio/Control Women dataset in 2000-2002. This sample is composed of a sufficient number of AN cases, parents of cases, and unrelated control women, to enable both the case:control and trio designs to be used in candidate gene association studies motivated by linkage and association findings identified in the Price Foundation AN-ARP dataset.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of the Price Foundation.

REFERENCES


