Published in final edited form as:

Mol Genet Metab. 2010 March ; 99(3): 275–282. doi:10.1016/j.ymgme.2009.10.189.

***In silico* and functional studies of the regulation of the glucocerebrosidase gene**

Yotam N. Blech-Hermoni^{*,1}, Shira G. Ziegler^{*,1}, Kathleen S. Hruska^{*,1}, Barbara K. Stubblefield¹, Mary E. LaMarca¹, Matthew E. Portnoy², NISC Comparative Sequencing Program³, Eric D. Green^{2,3}, and Ellen Sidransky^{1,+}

¹Section on Molecular Neurogenetics, Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892

²Physical Mapping Section, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892

³National Institutes of Health Intramural Sequencing Center (NISC), National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892

Abstract

In Gaucher disease (GD), the inherited deficiency of glucocerebrosidase results in the accumulation of glucocerebroside within lysosomes. Although almost 300 mutations in the glucocerebrosidase gene (*GBA*) have been identified, the ability to predict phenotype from genotype is quite limited. In this study, we sought to examine potential *GBA* transcriptional regulatory elements for variants that contribute to phenotypic diversity. Specifically, we generated the genomic sequence for the orthologous genomic region (~39.4 kb) encompassing *GBA* in eight non-human mammals. Computational comparisons of the resulting sequences, using human sequence as the reference, allowed the identification of multi-species conserved sequences (MCSs). Further analyses predicted the presence of two putative clusters of transcriptional regulatory elements upstream and downstream of *GBA*, containing five and three transcription factor-binding sites (TFBSs), respectively. A firefly luciferase (Fluc) reporter construct containing sequence flanking the *GBA* gene was used to test the functional consequences of altering these conserved sequences. The predicted TFBSs were individually altered by targeted mutagenesis, resulting in enhanced Fluc expression for one site and decreased expression for seven others sites. Gel-shift assays confirmed the loss of nuclear-protein binding for several of the mutated constructs. These identified conserved non-coding sequences flanking *GBA* could play a role in the transcriptional regulation of the gene contributing to the complexity underlying the phenotypic diversity seen in GD.

Keywords

Multi-species sequence comparisons; glucocerebrosidase; Gaucher disease; transcriptional regulation; luciferase assays

+Corresponding author: Section on Molecular Neurogenetics Medical Genetics Branch National Human Genome Research Institute Building 35, Room 1A213 35 Convent Drive, MSC 3708 Bethesda, MD 20892-3708 Phone: 301-451-0901 or 301-496-0373 Fax: 301-402-6438 sidranse@mail.nih.gov.

*These authors contributed equally

1. INTRODUCTION

Gaucher disease (GD) is caused by the dysfunction of a constitutive enzyme, glucocerebrosidase (GC), which cleaves a glucose moiety from the lipid glucocerebroside. Considerable clinical heterogeneity is encountered among patients with GD [1]. Associated GD phenotypes range from asymptomatic octogenarians to death *in utero*, and are grouped into three disease types based upon the presence and rate of progression of neurologic manifestations: Type 1 is also referred to as “nonneuronopathic GD”; type 2 as “acute neuronopathic GD”; and type 3 as “chronic neuronopathic GD”. However because of the spectrum of symptoms encountered, it is more accurate to regard GD as a continuum of phenotypes [2,3].

The phenotypic variability in GD raises questions about the consequences of glucocerebrosidase dysfunction. The disease is characterized by substrate accumulation and storage primarily in cells of the reticulo-endothelial lineage, giving rise to characteristic appearing Gaucher cells. GD affects the liver, spleen, bone and, in some cases, the brain. The stored lipid, glucocerebroside derives from the turnover of cell membranes; however, the process by which the lipid accumulation leads to the disease manifestations is largely unknown.

Almost 300 different mutations have been identified in patients with GD [4]. These mutations are scattered through the glucocerebrosidase gene (*GBA*), and include missense, nonsense, and frame-shift mutations as well as insertions, deletions, and complex alleles. The 11-exon *GBA* gene resides in a gene-rich region of human chromosome in1q21-22 [5]. A tandem duplication in the region is believed to have given rise to two unprocessed pseudogenes with both exonic and intronic sequence: a highly homologous pseudogene of *GBA* (*GBAP*); found 16 kb downstream of *GBA* [6] and pseudo-metaxin 1 (*MTX1P*), a pseudogene of the nearest downstream neighbor of *GBA*, metaxin 1 (*MTX1*). *GBA* and *GBAP* are both on one strand, while *MTX1* and *MTX1P* are found on the reverse strand (Figure 1). This duplication is present in the rhesus macaque, and presumably occurred before the divergence of the great apes and the Old World monkeys (25-40 Mya). Therefore much sequence similarity is retained between the functional genes and the corresponding pseudogenes. For example, *GBA* and *GBAP* share 96% sequence identity within the exons, with the latter lacking a 55-bp segment in exon 9. Some GD mutations appear to derive from the closely related *GBAP* pseudogene. In general, recombination events within and around *GBA* are relatively common [7], an observation confirmed by HapMap project data [8], and contribute to the complexity associated with genotypic analysis of *GBA*.

Numerous genotype/phenotype studies of GD demonstrate significant genotypic heterogeneity among clinically similar patients. Further, the vastly different phenotypes encountered among patients with GD, as well as among patients with other “monogenic disorders”, cannot be fully predicted by genotype [9,10]. Although some generalizations can be made, there is most often a range of phenotypes associated with any given mutation. Changes in tissue- or developmental stage-specific regulation of the *GBA* gene may influence disease severity. Moreover, the presence of *GBA* within a gene-rich genomic segment raises the potential for interactions with nearby genes or their regulatory elements.

Advances in comparative genome sequencing and analysis have yielded powerful new methods for identifying functional regions of the human genome, in particular non-coding sequences involved in regulating the expression of genes. Such an approach can be applied in a targeted fashion to study specific genes or genomic regions [11]. Computational methods have been developed to identify candidate regulatory regions based on the premise that functional non-coding sequences will have diverged significantly less than non-functional sequences. By comparing sequences from multiple species, sufficient statistical power is attained for

predicting with high confidence small blocks of sequence that are under strong evolutionary selection. These can then be further analyzed for the presence of putative transcription factor binding sites (TFBSs), with the latter then further studied using live cell reporter expression assays.

To maintain consistency with the literature, and to describe the relative location of the regulatory sites in the physical architecture of the gene region, “proximal promoter” and “distal promoter” refer to regions less than 1-kb and more than 1-kb of the +1 position, respectively. The term “flanking region” describes the combined upstream (up to about 5-kb of the start ATG) and downstream (up to about 3-kb of the stop TGA) regions surrounding the *GBA* gene.

Previous investigations of *GBA* regulation showed that both mRNA levels and enzyme activities vary among cell types, implicating regulation at both the transcriptional and protein levels [12]. Analyses of the human proximal promoter point to the presence of both regulatory enhancers and suppressors between nucleotide positions –354 and +267 of the beginning of exon 1 [13]. Some of these regulatory regions function constitutively, while others appear to function in a cell-specific fashion [13]. Gel-shift and DNase I footprinting assays of a more restricted region (positions –106 to +82 of the beginning of exon 1) demonstrate binding of transcription factors to some of the previously identified regulatory regions; protein binding was confirmed in two regions, while a third protein-binding site corresponded to a region previously reported to not affect reporter gene expression levels [14]. These discrepancies may be explained by the different cell types examined in the different studies.

Here, we generated and compared sequence of the genomic region encompassing *GBA* in multiple vertebrate species. The resulting analyses revealed highly conserved regions, which, in turn, were further characterized for their role in regulating *GBA* expression.

2. MATERIALS AND METHODS

2.1 Acquisition and Analysis of Sequence Data

Using probes from small stretches of sequence conserved between human and mouse, the NIH Intramural Sequencing Center (NISC) isolated and sequenced bacterial artificial chromosomes (BAC) clones containing the *GBA* locus from 7 mammals (galago, cow, dog, hedgehog, armadillo, fruit bat and wallaby). The acquired sequences were published in Genbank both as individual BAC clones and as non-redundant multi-BAC assemblies for each species, with the following accession numbers for the multi-BAC assemblies: armadillo-T99 DP000335; cow-T99 DP000332; bat-T99 DP000333; dog-T99 DP000334; galago-T99 DP000335; hedgehog-T99 DP000336; and wallaby-T99 DP000337. The direct link to these assemblies is at <http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Search&db=nucleotide&term=DP000>. Orthologous genomic DNA sequence from chimpanzee and mouse were obtained online from recent sequence assemblies at NCBI and Ensembl.

Draft sequence assemblies from the BAC clones and current sequences from the other organisms were aligned using MultiPipMaker (<http://pipmaker.bx.psu.edu/pipmaker>). Multi-species Conserved Sequences (MCSs) were identified and mapped with WebMCS (<http://research.nhgri.nih.gov/MCS>) and the UCSC Genome Browser (<http://www.genome.ucsc.edu>). For these alignments, a DNA segment spanning 39.4 kb from the beginning of *THBS3* through the end of *C1orf2* at the *GBA* locus was imported. While a significantly larger region of sequence at 1q21-1q22 (177.6kb) was provided by NISC, the analysis reported herein was limited to exclude *THBS3* because of the high homology observed in exonic sequences in initial MCS analyses.

These sequences were also analyzed with the applications available through the GenomatixSuite (<http://www.genomatix.de>). FrameWorker was utilized to identify orientation- and distance-defined modules of transcription factor binding sites.

2.2 Construct Design and Site-Directed Mutagenesis

An expression construct including the conserved regions identified by both *in silico* approaches was produced. Using a PCR-based insertion strategy [15], 8.1 kb of the human *GBA* (hGBA) locus from BAC clone RP11-263K19 was subcloned sequentially into a firefly luciferase reporter vector (pGL4.10, Promega Corporation)(Supplemental Data, Table 1a). The construct was made in three sequential cloning steps (pGL4.10-hGBApro, -hGBApro+3'UTR, and -hGBAloci). The promoter region (hGBApro) was initially cloned in, followed by the 3'-UTR. First, 2.3 kb of the promoter region, including genomic sequence through the first six codons of *GBA* exon 3, were cloned in-frame to the fourth codon of the firefly luciferase cDNA. This yielded an expression construct (pGL4.10-hGBApro) with both initiation codons of *GBA* and the full-length signal peptide. This construct was tested extensively to confirm that the resulting transcript was spliced appropriately and inclusion of the lysosomal signal peptide did not destroy luciferase activity (data not shown). Next, 2.9 kb of the 3'-UTR through intron 3 of pseudo-MTX1 was cloned downstream of the luciferase cDNA (pGL4.10-hGBApro+3UTR). An additional 2.9kb was added upstream of the promoter to create the final pGL4.10-hGBAloci construct.

The resulting construct was altered by site-directed mutagenesis to individually disrupt the TFBSs (Supplemental Data, Table 1b). SequenceShaper, another application available in the Genomatix Suite, was utilized to determine 2-bp substitutions that would disrupt transcription factor binding without creating potential sites for additional factors. Eight constructs (V\$CLOX, V\$ETSF2, V\$MITF, V\$EBOX, V\$ETSF3, V\$CMYB, V\$NF1F, V\$MOKF) were made using the QuikChange Site-Directed Mutagenesis kit (Stratagene). All constructs were transformed into TOP-10 chemically competent cells and confirmed by sequencing.

2.3 Transfections

Each mutated construct was transfected independently into COS-7 cells and the resulting luciferase expression was measured. Individual transfections were performed with 94 replicates and every construct was repeated two to five times (see *Statistical Analysis*, below for determination of minimum corrected sample size). The wildtype pGL4.10-hGBAloci control construct was included in every transfection experiment. Briefly, cells were seeded in 96-well plates at approximately 8×10^3 cells per well. At 90% confluency, media was changed to antibiotic-free media and cells were then co-transfected with each of the Fluc reporter constructs and a Renilla luciferase (Rluc) control construct (phRL-TK, Promega Corporation). For each plate, 36.3µl of FuGENE6 Transfection Reagent (Roche) was diluted in Opti-MEM Medium (GIBCO) to a volume of 605µl and allowed to incubate for 5 min at RT. 60.5µl (0.2ug/µl) DNA solution (0.196 µg/µl pGL4.10-hGBAloci construct + 0.004µg/µl phRL-TK) was added to the FuGENE6:Opti-MEM mixture, mixed gently, and incubated for 40 min at RT, resulting in a reagent:DNA ratio of 6:2. Using a multi-channel pipette, 5.5ul of DNA:FuGENE6:Opti-MEM was added in each well. Two wells on each plate were left untransfected to serve as blank controls in the dual luciferase assay. Transfected cells were incubated at 37°C for 48 hours.

2.4 Dual-Luciferase Assay

Following the incubation period, cells were washed with 100µl 1X Puck's Saline, lysed with 20µl 1X Passive Lysis Buffer (Dual-Luciferase Reporter 1000 Assay System, Promega Corporation), and subjected to two freeze-thaw cycles (frozen on dry-ice and thawed at room temperature). The cell lysates then were assayed using the Dual-Luciferase Reporter 1000

Assay System (Promega), according to the manufacturer's instructions. Activities of firefly and Renilla luciferase were measured sequentially from each well using a VICTOR multilabel plate reader (Victor³V, PerkinElmer). Data were expressed as ratio of firefly/Renilla luciferase (Fluc/Rluc) activity to control for transfection efficiency.

2.5 Statistical Analyses

The statistical analyses were performed on the dual luciferase assay data as described by Jacobs and Dinman [16]; <http://dinmanlab.umd.edu/statistics/>) for a continuous, ratiometric data set. The statistics were calculated from sets of ratiometric luminescence values taken from multiple experiments and cell lysates. For each data point, the relative ratio of firefly luciferase luminescence to Renilla luminescence was calculated. These data then were pooled for each construct and the usual descriptive statistics calculated (sample mean, median, variance, standard deviation and standard error of the sample mean) (Supplemental Data, Table 2, Figure 5). Outliers were identified and excluded based on upper and lower boundaries of the median $\pm (1.5 \times (75^{\text{th}} \text{ centile} - 25^{\text{th}} \text{ centile}))$. The minimum uncorrected sample size ($\tilde{N}$) was calculated for a 0.95 confidence level (i.e., $\alpha = 0.05$) and the minimum corrected sample size (N^*) was determined from a reference table (available at <http://dinmanlab.umd.edu/statistics/>). Probability plots were constructed (data not shown) and the corresponding normal probability plot correlation coefficients (PPCCs) and critical values for rejection at the 0.95 confidence level were determined for each set of dual luciferase data. Once each ratiometric data set was trimmed of outliers, passed the test for normality and was found to be of sufficient size, the ratiometric gene expression of each experimental reporter vector relative to that of the control, wild-type *GBA* reporter was calculated (Supplemental Data, Table 2).

2.6 Nuclear protein extraction

The extraction of nuclear proteins was modified from that used by Moran et al. [14]. Briefly, Two T75 flasks of adherent cells (COS-7; ~80% confluence) were trypsinized and collected by centrifugation (5 min at 1,200 rpm on a bench-top Sorvall T6000D). Pellets were washed in 30 volumes of ice-cold PBS and repelleted. Pellets were resuspended in 1 volume of Buffer A (10mM HEPES, 1.5mM MgCl₂, 10mM KCl, 0.5mM DTT) and incubated on ice for 15 min. Cells were lysed by passing through a 25G needle 5 times, and samples were combined. Nuclei were pelleted by centrifugation (20 sec at 12,000g on an Eppendorf 5415D microcentrifuge). The pellet was resuspended in 2 original volumes of Buffer C (20mM HEPES, 25% glycerol, 420mM NaCl 1.5mM MgCl₂, 0.2mM EDTA, 0.5mM DTT, protease inhibitor cocktail tablet (Roche Applied Science, Indianapolis, IN). The suspension was incubated on ice for 30 min with vigorous shaking every 5 min. Debris was pelleted by centrifugation at 12,000g for 5 min. The supernatant was dialyzed against Buffer D (20mM HEPES, 20% glycerol, 0.1M KCl, 0.2mM EDTA, 0.5mM DTT, protease inhibitor cocktail tablet) for 2 hrs (MINI Slide-alyzer; Thermo Scientific, Rockford, IL [formerly Pierce Biotech]). Samples were aliquoted, immediately flash-frozen on EtOH/dry-ice, and stored at -80°C until needed.

2.7 Probe preparation

Probe was prepared from separately ordered oligonucleotides (29-31nt; Bioserve Biotechnologies, Gaithersburg, MD). Lyophilized oligos were resuspended to 100 μM in dilution buffer (10mM Tris [pH 8.0], 1mM EDTA, 50mM NaCl), annealed as described below, and diluted to final concentrations (63nM for probe and 50 μM for competitor) in water. Probe was prepared by annealing a labeled oligo (5' biotinylated) to a complementary (unlabeled) oligo by combining the two oligos to a concentration of 50 μM , heating to 95°C for 5 min in a water bath, and allowing to cool naturally. "Cold" competitor was created identically, by using two unlabeled oligos.

Probe and competitor sequences were identical to the primers used for construct mutagenesis (described above; shown in Table 1b).

2.8 Electrophoretic Mobility Shift Assays (EMSA)

Binding reactions (20 μ L total volume) were carried out using the LightShift Chemiluminescent EMSA kit (Thermo Scientific), as described by the manufacturer. Binding solution (2 μ L binding buffer, 1 μ L glycerol, 1 μ L MgCl₂, 1 μ L Poly(dI-dC), water) was prepared on ice. Probe (126fmol) was added (and, in competition assays, 100pmol competitor was added). Nuclear protein extract (10 μ g) was added last, the solutions were mixed gently, and the reactions were incubated on the bench for 30 min. Concurrently, 6% acrylamide DNA retardation gels (Invitrogen Corporation, Carlsbad, CA) were prerun at 100V for 30 min (in 0.5X TBE running buffer). After incubation, 5 μ L loading buffer was added to each binding reaction and the whole sample was loaded onto the gel. Gels were run at 100V for 1 hr. Samples were transferred from the gel to charged Nitrocellulose (Hybond N+; Amersham) membranes using the iBlot semi-dry transfer apparatus (Invitrogen). Blots were detected using the Chemiluminescent Nucleic Acid Detection Module (Pierce). Blocking and Wash Buffers were preheated to 50°C. All incubations were carried out in 50 cc conical tubes rotating in a hybridization oven at room temperature. Membranes were first blocked in 5 mL Blocking Buffer for 15 min. Blocking Buffer was then exchanged with 5 mL HRP-conjugate (1:50,000 in Blocking Buffer), and followed by 5 min incubation. Membranes were transferred to clean tubes and washed 5 times in 5 mL 1X Wash Buffer. Membranes were again transferred to clean tubes and washed in 10 mL Substrate Equilibration Buffer. Membranes were blotted briefly to remove excess liquid, bathed in 3 mL detection solution for 5 min, and immediately visualized by exposure to autoradiography film (1-5 min).

3. RESULTS

We performed comparative mapping and sequencing of the genomic region encompassing *GBA* in 8 vertebrate species, generating a total of 2.25 Mb of high-quality genomic sequence data. The resulting sequences were aligned using the available human sequence as the reference. We used WebMCS (formerly <http://zoo.nhgri.nih.gov/mcs/>; [17]) to identify multi-species conserved sequences (MCSs) within the multi-sequence alignments, focusing our attention on the non-coding (intronic and inter-genic) regions. This identified a known upstream enhancer of thrombospondin 3 (*THBS3*), which is located in intron 6 of the *MTX1* [18], and a second (novel) MCS was detected in intron 3 of the *MTX1*. These sequences are also found in *MTX1P* (data not shown).

Further analyses using the FrameWorker application (part of the Genomatix Suite) identified two clusters of TFBSs, (Figure 2, designated as “upstream module” and “downstream module”). The upstream module, located within 250 bp of the first initiation codon of *GBA*, consists of five TFBSs and is not present in *GBAP*. The downstream module, which spans from the 3'-UTR of *GBA* through intron 3 of *MTX1*, consists of three TFBSs. Both modules are highly conserved in the analyzed sequences, with two notable exceptions: The upstream module is incomplete in the primate *GBAP* and both modules are incomplete in the mouse sequence.

The Genomatix suite provides an *in silico* tool that predicts nucleotide changes that would abolish protein binding to each TFBS, without altering predicted binding to neighboring sites. Figure 3 shows the multi-sequence alignment, with conserved positions denoted by dots, and less conserved positions denoted by the nucleotide appearing at that location in each species. In addition, the critical regions for the eight TFBSs are designated by boxed nucleotides, and red letters above the alignment designate the locations and alteration that is predicted to abolish binding at each binding site. Transcription factor family definitions were obtained from

Genomatix and from TraFaC (Transcription Factor Binding Site Comparison; developed by the Biomedical Informatics Division of Cincinnati Children's Hospital Medical Center): The upstream module contains predicted binding sites for five transcription factor families: V\$CLOX includes CLOX and CLOX-homology factors, as well as CUT-like domain transcriptional repressor (CDP); V\$ETSF1 and V\$ETSF2 include human and murine ETS1 factors, such as c-Ets-1 (p54), Elk-1, and NERF1a; V\$MITF includes Microphthalmia Transcription Factor (MITF) and TFE3; and V\$EBOX includes E-box binding transcription factors, such as cMyc/MAX heterodimers and nMyc. The downstream module contains predicted binding sites for three transcription factor families: V\$MOKF includes mouse Kruppel-like factors and human MOK-2; V\$NF1F includes Nuclear Factor 1; and V\$CMYB includes cMyb, a cellular transcription activator (http://www.genomatix.de/online_help/help_gems/mat_lib_30.html).

The genomic regions upstream and downstream of human *GBA* were subcloned into the promoterless firefly luciferase (Fluc) reporter construct pGL4.10 in a step-wise fashion (Figure 4). Eight different reporter constructs were created, each with a different TFBS abolished. For each construct, two nucleotides were mutated, corresponding to the red-labeled nucleotides in Figure 3. In order to normalize expression values between transfected wells, each well was co-transfected with a Fluc construct (wild-type or mutant), as well as with a control plasmid for basal expression of *Renilla* luciferase (Rluc). Assays for each construct were carried out in a 96-well format, producing 94 experimental replicates, and two control wells with untransfected cells, per assay. Both firefly and *Renilla* luciferase activities were measured in each well, and the data are expressed as ratios of firefly/*Renilla* luciferase (Fluc/Rluc) activity to control for transfection efficiency. In COS-7 cells, the wild-type construct produced an expression level of 15.06 Relative Luminescence Units (RLU), while all alterations to the native sequence changed this level of expression (Figure 5). One construct, delV\$ETSF2, in which the first V\$ETSF site in the upstream module was eliminated, resulted in a 29.2% increase in expression (19.46 RLU). All other alterations to the predicted binding sites resulted in a robust decrease in expression (between 3.34 RLU and 0.90 RLU, or 77.8% to 94% of wild-type). Although all changes with respect to wild-type achieved statistical significance, these data do not necessarily reflect biological significance of such increases or decreases in protein expression levels.

Using probes for the wild-type and mutant downstream V\$MOKF and V\$CMYB sites, gel-shift assays were carried out to confirm loss of binding to the predicted sequence alterations. In Figure 6, two assays are presented, in which a biotin-labeled oligonucleotide probe is exposed to increasing amounts of nuclear protein extract, and then passed through a polyacrylamide gel. On the left side of each panel, a positive control (wtV\$MOKF; a probe representing wild-type sequence of the V\$MOKF TFBS) demonstrates that in the absence of protein, the oligonucleotide travels to the bottom of the gel, while after exposure to nuclear protein extract (10ug) the bound complex travels as a much larger unit through the gel. On the right, similar amounts of probe (126fmol) were exposed to increasing amounts of nuclear extract (from 0ug to either 10ug or 50ug): In the top panel, even at excess amounts of protein, no binding can be seen to the mutated V\$MOKF probe; in the bottom panel, binding is seen to the wild-type V\$CMYB probe (even at 2.5ug protein), but not the mutated V\$CMYB (Figure 6).

4. DISCUSSION

The vast clinical heterogeneity observed among patients with Gaucher disease as well as other "simple" Mendelian disorders has not been adequately explained by focusing on the role of specific disease causing mutant alleles. While the gene encoding glucocerebrosidase has been known and characterized for over two decades, our understanding of its regulation is far less

well understood. The current availability of comparative genomic sequence data provides a new tool to revisit this issue and to explore regions involved in gene regulation of *GBA*.

In this study high quality sequence was generated from eight mammalian species and the genomic and flanking regions of *GBA* in these mammals as well as sequence from human, chimpanzee and mouse were evaluated. The value of comparing gene structure among different species is three fold: it enables us to explore evolutionary relationships, it enhances gene prediction algorithms, and it permits us to identify functionally constrained sequence. The basic premise is that mutations in functional sequences are less likely to be tolerated than those in neutrally-evolving sequences. We set out to identify sequences that have diverged less than expected assuming that these sequences are more likely to have a functional role. In this study we initially looked for MCSs, which are sequences accounting for only 5% of genomic sequence [19]. Unfortunately, for *GBA* this approach was not extremely fruitful, yielding only two MCSs in the region, one of which was a known enhancer for another nearby gene. We then turned to other *in silico* methods to identify regulatory modules, looking for predicted transcription factor binding sites based upon evolutionary conservation. Here we identified two novel regulatory regions that were explored using *in vitro* strategies. Our results indicate that these regions did indeed have functional significance, further validating the role of comparative genomic strategies.

Our data provide functional confirmation of the predictive power offered by computational methods, both in identifying multi-species conserved regulatory sites and in estimating the effects of nucleotide alterations on the integrity of predicted binding sites. Ultimately, our goal will be to use these computational methods as a tool for the detection of sites important to the regulation of a disease-causing gene product that could impact individual patients. The deficiency of an activator of GC, saposin C, has previously been shown to give rise to a phenotype resembling GD [20-22]. This work identifies regulatory variants that could similarly impact GD phenotype through GC deficiency unrelated to the mutations in the coding region.

Two approaches have dominated the study of gene regulation. One has been to functionally query flanking or non-coding sequence of candidate genes within a particular tissue or cell-type. The other has come of the great recent advances in bioinformatics algorithms, enabling us to probe sequences or sequence alignments for consensus binding sites or evolutionarily conserved sequences. While the computational power of bioinformatics approaches is unrivaled, comprehensive validation is the rate-limiting step in the pipeline, resulting in massive efforts focusing on site discovery and little by way of functional confirmation. Thus, while this methodology can identify sites that have been conserved over millions of years of evolution, it may not be able to resolve clade-specific (i.e., human- or primate-specific) regulatory elements. By contrast, gene-centric approaches provide a restricted perspective of regulation by focusing on active regulatory elements in a particular cell type or disease-specific context.

While the computational algorithms used to identify multi-species conserved regulatory modules in this study failed to identify other published regulatory sites, we did find overlap with previously described regulatory regions. Doll *et al.* [13] identified several “downstream stimulatory regions” (DSRs) by destroying portions of the proposed promoter, three of which (DSR3, DSR4, and DSR5) overlap with our described “Upstream Module.” However, their work did not precisely indicate binding sites, but rather described regions which they considered “stimulating”, representing the sum of the combinatorial effect of any number of positive and negative regulators of transcription. Although our approach did not identify the specific binding sites described by Moran *et al.* [14], there was also juxtaposition of their sites and our “Upstream Module.” In fact, overlap exists between the V\$EBOX, described here, and an AP-1 binding site described by Moran *et al.* as well as, to a lesser extent, with neighboring

sites, such as V\$MITF and V\$ETSF1..Both of these previous reports, investigated the activity of regulatory regions, or elements, only in specific cells, or under particular conditions. This limitation was recognized by Moran *et al.* in their discussion of the possible role of one of their OCT binding motifs, known to be differentially utilized [14].

Our study explored the predictive potential of a computational method to investigate the impact of multi-species conserved sequences on the regulation of a gene responsible for a Mendelian disorder, GD. With the rapidly growing predictive power of computational tools, these techniques will soon be applied to genes associated with complex phenotypes. As the resolution of regulatory elements from previously indecipherable sequence becomes tractable, their potential for cell-, tissue-, and developmental stage-specific modulation can finally be investigated proactively. Insights from studies in Mendelian disorders may ultimately prove useful in designing strategies for functional evaluations of the many conserved regulatory regions identified throughout the human genome.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Intramural Research Program of the National Human Genome Research Institute and National Institutes of Health.

REFERENCES

- [1]. Beutler, E.; Grabowski, G. Gaucher Disease. In: Scriver, CBA.; Sly, W.; Valle, D., editors. The Metabolic & Molecular Bases of Inherited Disease. McGraw-Hill; New York: 2001. p. 3635-3668.
- [2]. Sidransky E. Gaucher disease: complexity in a "simple" disorder. Molecular genetics and metabolism 2004;83:6–15. [PubMed: 15464415]
- [3]. Goker-Alpan O, Schiffmann R, Park JK, et al. Phenotypic continuum in neuronopathic Gaucher disease: an intermediate phenotype between type 2 and type 3. The Journal of pediatrics 2003;143:273–276. [PubMed: 12970647]
- [4]. Hruska KS, LaMarca ME, Scott CR, et al. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Human mutation 2008;29:567–583. [PubMed: 18338393]
- [5]. Winfield SL, Tayebi N, Martin BM, et al. Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. Genome research 1997;7:1020–1026. [PubMed: 9331372]
- [6]. Horowitz M, Wilder S, Horowitz Z, et al. The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics 1989;4:87–96. [PubMed: 2914709]
- [7]. Tayebi N, Stubblefield BK, Park JK, et al. Reciprocal and nonreciprocal recombination at the glucocerebrosidase gene region: implications for complexity in Gaucher disease. American journal of human genetics 2003;72:519–534. [PubMed: 12587096]
- [8]. Frazer KA, Ballinger DG, Cox DR, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–861. [PubMed: 17943122]
- [9]. Koprivica V, Stone DL, Park JK, et al. Analysis and classification of 304 mutant alleles in patients with type 1 and type 3 Gaucher disease. American journal of human genetics 2000;66:1777–1786. [PubMed: 10796875]
- [10]. Dipple KM, McCabe ER. Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. American journal of human genetics 2000;66:1729–1735. [PubMed: 10793008]
- [11]. Thomas JW, Touchman JW, Blakesley RW, et al. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 2003;424:788–793. [PubMed: 12917688]

- [12]. Doll RF, Smith FI. Regulation of expression of the gene encoding human acid beta-glucosidase in different cell types. *Gene* 1993;127:255–260. [PubMed: 8500768]
- [13]. Doll RF, Bruce A, Smith FI. Regulation of the human acid beta-glucosidase promoter in multiple cell types. *Biochimica et biophysica acta* 1995;1261:57–67. [PubMed: 7893761]
- [14]. Moran D, Galperin E, Horowitz M. Identification of factors regulating the expression of the human glucocerebrosidase gene. *Gene* 1997;194:201–213. [PubMed: 9272862]
- [15]. Lee J, Lee HJ, Shin MK, et al. Versatile PCR-mediated insertion or deletion mutagenesis. *BioTechniques* 2004;36:398–400. [PubMed: 15038153]
- [16]. Jacobs JL, Dinman JD. Systematic analysis of bicistronic reporter assay data. *Nucleic Acids Res* 2004;32:e160. [PubMed: 15561995]
- [17]. Margulies EH, Blanchette M, Haussler D, et al. Identification and characterization of multi-species conserved sequences. *Genome research* 2003;13:2507–2518. [PubMed: 14656959]
- [18]. Collins M, Rojnuckarin P, Zhu YH, et al. A far upstream, cell type-specific enhancer of the mouse thrombospondin 3 gene is located within intron 6 of the adjacent metaxin gene. *J Biol Chem* 1998;273:21816–21824. [PubMed: 9705320]
- [19]. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;420:520–562. [PubMed: 12466850]
- [20]. Christomanou H, Chabas A, Pampols T, et al. Activator protein deficient Gaucher's disease. A second patient with the newly identified lipid storage disorder. *Klin Wochenschr* 1989;67:999–1003. [PubMed: 2615292]
- [21]. Pampols T, Pineda M, Giros ML, et al. Neuronopathic juvenile glucosylceramidosis due to sap-C deficiency: clinical course, neuropathology and brain lipid composition in this Gaucher disease variant. *Acta Neuropathol* 1999;97:91–97. [PubMed: 9930900]
- [22]. Diaz-Font A, Cormand B, Santamaria R, et al. A mutation within the saposin D domain in a Gaucher disease patient with normal glucocerebrosidase activity. *Hum Genet* 2005;117:275–277. [PubMed: 15856305]

Figure 1.

The *GBA* locus. Local topology of the genomic sequence surrounding the *GBA* gene (in humans). *GBA* and *GBAP* are both on the reverse strand, antiparallel to both *MTXI* and *MTXIP*. Following the early primate duplication, *MTXIP* in primates occupies the spot previously held by *MTXI* (earlier in evolution). Coordinates reflect NCBI Build 36.1 (hg18).

Upstream Module

Downstream Module

Figure 2. Predicted transcription factor-binding sites (TFBSs) identified by Genomatix Frameworker. The “Upstream Module” (*left*) and the “Downstream Module” (*right*): a cartoon shows the location and orientation of different predicted TFBSs, and a schematic representation of the genomic region including exons 1 and 2 of *GBA* (upstream) and exons 3 and 4 of *MTX1/MTX1P* (downstream) illustrate the relative location and composition of analogous clusters identified in various species and human pseudogene.

Figure 3. Multi-sequence alignment of the *GBA* promoter region (A) and intron 3 of the *MTX1P* (B). The specific TFBSs and the nucleotides altered are annotated.

Figure 4.

The design of the luciferase construct (pGL4.10-hGBAloci) is shown. Using a PCR-based insertion strategy, 8.1 kb of the *GBA* locus encompassing both upstream and downstream sequence was subcloned stepwise into pGL4.10.

Figure 5.

Luciferase expression in transfected COS-7 cells. Luciferase expression, as represented as firefly luciferase (Fluc) normalized to *Renilla* luciferase (Rluc), is shown for eight transfected reporter constructs. Five upstream and three downstream mutated TFBSs are shown. Inserts show a schematic representation of the upstream and downstream modules, showing the strand orientation of the sites.

Figure 6.

Gel-shift assays were performed to validate the loss of binding to the predicted sequence alterations in the TFBSs. For the V\$MOKF (A) and V\$CMYB (B) sites, protein binding to the wild-type sequence was confirmed, while a loss of binding was observed with the mutant sequence. Probe concentrations (126 fmol) are the same in all lanes; unless stated otherwise, 10 μg protein extract was used.

Table 1a

Primers used in cloning constructs.

hGBApro(a2)	TCGGCACCCGCACTACACAATTCTCTTAC
hGBApro(b1)	GCCCAGGCTGCTTTACCAACAGTACCG
hGBApro(c1)	TGGTAAAGCAGCCTGGGCAACAGAGTGAG
hGBApro(d1)	GTTCTTGGCGATGCAGGGTCGGGCACC
hGBApro(e1)	CCCTGCATCGCCAAGAACATTAAGAAG
hGBApro(f1)	CGCTCGTTGTAGATGTCGTTAGCTGG
hGBA3UTR(a1)	GTCATCCCCGACACCGCTATCCTCAGC
hGBA3UTR(b1)	GCAAGATCGCCGTGTAGATGGAGCAG
hGBA3UTR(c1)	TCGCCGTGTGATGGAGCAGATACTCAAG
hGBA3UTR(d1)	CTTCCCGACCCAGACACCCTTGAGAG
hGBA3UTR(e1)	CCCAGACACCTTGAGAGCCTTCAACCC
hGBA3UTR(f1)	GTGTAGGTCGTTTCGCTCCAAGCTGGTG
hGBAloci(g1)	GAGGAAGCTTATACTAAAGTGGCAGGGTGG
hGBAloci(h1)	TGACACAGGAAGTGAGGCAATCACAG

Table 1b

Primers used in construct mutagenesis.

V\$CLOX	5'	GAA	GCC	AAT	TGC	CCT	TTA	GAA	ATC	AGG	CTG	TGA	TTG	CC	3'
	3'	CTT	CGG	TTA	ACG	GGA	AAT	CTT	TAG	TCC	GAC	ACT	AAC	GG	5'
V\$ETSF2	5'	GAG	CGT	CAC	ATG	ACA	CAG	AGA	GTG	AGG	CAA	TCA	CAG		3'
	3'	CTC	GCA	GTG	TAC	TGT	GTC	TCT	CAC	TCC	GTT	AGT	GTC		5'
V\$MITF	5'	GCC	TCA	CTT	CCT	GTG	TAC	TGT	GAC	GCT	CCT	AGT	C		3'
	3'	CGG	AGT	GAA	GGA	CAC	ATG	ACA	CTG	CGA	GGA	TCA	G		5'
mV\$MITF	5'	GCC	TCA	CTT	CCT	GCA	TCA	CGT	GAC	GCT	CCT	AGT	CAT	C	3'
	3'	CGG	AGT	GAA	GGA	CGT	AGT	GCA	CTG	CGA	GGA	TCA	GTA	G	5'
V\$EBOX	5'	GTG	ACG	CTC	CTA	GTC	ATA	CCA	TGA	CCC	ATC	CAC	ATC		3'
	3'	CAC	TGC	GAG	GAT	CAG	TAT	GGT	ACT	GGG	TAG	GTG	TAG		5'
mV\$EBOX	5'	CAT	GTG	ACG	CTC	CTA	GTC	ATC	ATG	TGA	CCC	ATC	CAC		3'
	3'	GTA	CAC	TGC	GAG	GAT	CAG	TAG	TAC	ACT	GGG	TAG	GTG		5'
V\$ETSFI	5'	CCA	TCC	ACA	TCG	GGA	AGC	CGC	CAT	TAC	TTG	CAG	GG		3'
	3'	GGT	AGG	TGT	AGC	CCT	TCG	GCG	GTA	ATG	AAC	GTG	CC		5'
V\$CMYB	5'	CCA	GGC	TTT	TTG	CCA	ACA	GAG	ACA	GGC	TAC	TGG	CTG		3'
	3'	GGT	CCG	AAA	AAC	GGT	TGT	CTC	TGT	CCG	ATG	ACC	GAC		5'
V\$NFIF	5'	CAA	GGC	CCA	GGC	TTT	TTT	CCG	ACA	GCA	ACA	GGC	TAC		3'
	3'	GTT	CCG	GGT	CCG	AAA	AAA	GGC	TGT	CGT	TGT	CCG	ATG		5'
V\$MOKF	5'	TTG	ACG	GCC	AGC	GGG	CTC	GAA	TGC	CCA	GGC	TTT	TTG		3'
	3'	AAC	TGC	CGG	TCG	CCC	GAG	CTT	ACG	GGT	CCG	AAA	AAC		5'
mV\$MOKF	5'	TTG	ACG	GCC	AGC	GGG	CTC	CAA	AGC	CCA	GGC	TTT	TTG		3'
	3'	AAC	TGC	CGG	TCG	CCC	GAG	GTT	TCG	GGT	CCG	AAA	AAC		5'
	5'	CAA	AGC	CCA	GGC	TTT	CTG	CCA	ACA	GCA	ACA	GGC	TAC		3'
	3'	GTT	TCG	GGT	CCG	AAA	GAC	GGT	TGT	CGT	TGT	CCG	ATG		5'

Table 2

COS-7 Cell Data		wt	AV\$CLOX	AV\$TSEF2	AV\$MITF	AV\$EBOX	AV\$TSEF3	AV\$CMYB	AV\$NFIF	AV\$MOKF
$Q_{\max}$		49.0880	3.8610	43.1172	5.3136	2.5005	8.1967	4.5629	6.6969	4.7301
Q_{75}		18.5741	2.3748	24.4281	2.3067	1.1996	4.1637	2.4218	3.0666	2.2633
$\approx x$		15.2066	1.9125	19.6238	1.8849	0.9177	3.3177	1.9165	2.3051	1.8620
Q_{25}		12.3168	1.2500	14.9767	1.4326	0.5962	2.7142	1.6299	1.6274	1.4700
$Q_{\min}$		5.6466	0.0442	6.9880	0.4091	0.1195	0.1961	0.8913	0.3429	0.1052
$Q_{75} - Q_{25}$		6.2573	1.1249	9.4513	0.8741	0.6034	1.4495	0.7920	1.4392	0.7932
O_U		24.5925	3.5998	33.8008	3.1960	1.8228	5.4919	3.1045	4.4640	3.0518
O_L		5.8207	0.2252	5.4468	0.5738	0.0126	1.1435	0.7286	0.1463	0.6722
Outliers?		34	5	9	13	14	29	22	15	21
x		15.0565	1.8407	19.4623	1.8331	0.8977	3.3386	1.9042	2.2596	1.8809
s^2_{N-1}		1.36×10^1	5.26×10^{-1}	3.32×10^1	2.82×10^{-1}	1.31×10^{-1}	8.43×10^{-1}	2.20×10^{-1}	8.34×10^{-1}	2.94×10^{-1}
s_{N-1}		3.68×10^0	7.25×10^{-1}	5.76×10^0	5.31×10^{-1}	3.61×10^{-1}	9.18×10^{-1}	4.69×10^{-1}	9.13×10^{-1}	5.42×10^{-1}
S_e		1.80×10^1	4.36×10^{-2}	4.30×10^{-1}	4.11×10^{-2}	2.21×10^{-2}	6.03×10^{-2}	3.64×10^{-2}	5.62×10^{-2}	4.20×10^{-2}
$\tilde{N}$		92	239	135	129	250	117	94	252	128
N^*		109	247	153	147	256	135	111	258	146
N		417	277	179	167	268	232	166	264	167
Sufficient Sampling?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
PPCC	0.9959	0.9956	0.9923	0.9948	0.9942	0.9942	0.9955	0.9890	0.9972	0.9916
Critical value	0.9951	0.9929	0.9887	0.9887	0.9924	0.9924	0.9914	0.9887	0.9924	0.9887
Normal distribution?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
x_R	--	0.1223	1.2926	0.1218	0.0596	0.0596	0.2217	0.1265	0.1501	0.1249
s^2_R	--	3.24×10^{-3}	3.25×10^{-2}	3.09×10^{-3}	1.63×10^{-3}	1.63×10^{-3}	4.81×10^{-3}	2.85×10^{-3}	4.14×10^{-3}	3.16×10^{-3}
s_R	--	1.16×10^{-1}	2.49×10^0	1.11×10^{-1}	5.44×10^{-2}	5.44×10^{-2}	2.08×10^{-1}	1.15×10^{-1}	1.46×10^{-1}	1.14×10^{-1}
$s_e(x_R)$	--	3.21×10^{-3}	2.46×10^{-1}	2.13×10^{-3}	7.89×10^{-4}	7.89×10^{-4}	6.66×10^{-3}	1.93×10^{-3}	5.03×10^{-3}	2.23×10^{-3}