A Brain Region-Specific Predictive Gene Map for Autism Derived by Profiling a Reference Gene Set

Ajay Kumar1, Catherine Croft Swanwick1, Nicole Johnson1, Idan Menashe1, Saumyendra N. Basu1, Michael E. Bales2, Sharmila Banerjee-Basu1*

Abstract

Molecular underpinnings of complex psychiatric disorders such as autism spectrum disorders (ASD) remain largely unresolved. Increasingly, structural variations in discrete chromosomal loci are implicated in ASD, expanding the search space for its disease etiology. We exploited the high genetic heterogeneity of ASD to derive a predictive map of candidate genes by an integrated bioinformatics approach. Using a reference set of 84 Rare and Syndromic ASD genes (AutRef84), we built a composite reference profile based on both functional and expression analyses. First, we created a functional profile of AutRef84 by performing Gene Ontology (GO) enrichment analysis which encompassed three main areas: 1) neurogenesis/projection, 2) cell adhesion, and 3) ion channel activity. Second, we constructed an expression profile of AutRef84 by conducting DAVID analysis which found enrichment in brain regions critical for sensory information processing (olfactory bulb, occipital lobe), executive function (prefrontal cortex), and hormone secretion (pituitary). Disease specificity of this dual AutRef84 profile was demonstrated by comparative analysis with control, diabetes, and non-specific gene sets. We then screened the human genome with the dual AutRef84 profile to derive a set of 460 potential ASD candidate genes. Importantly, the power of our predictive gene map was demonstrated by capturing 18 existing ASD-associated genes which were not part of the AutRef84 input dataset. The remaining 442 genes are entirely novel putative ASD risk genes. Together, we used a composite ASD reference profile to generate a predictive map of novel ASD candidate genes which should be prioritized for future research.

Introduction

Autism (MIM 209850) is a broad-spectrum multi-factorial condition which onsets in the first years of life and persists throughout the lifetime [1]. A triad of deficits in the areas of social communication, language development, and repetitive activities/restricted range of interests defines the core symptoms used in the diagnosis of autism (DSM IV, 1994). The affected areas show a broad range of variability in terms of both symptoms and severity; co-morbidity of epilepsy and mental retardation are often observed. Autism spectrum disorders (ASD) is a commonly used term to cover the wide variations of autism. The dramatic rise in the prevalence of ASD in recent years is of major public concern [2–3].

A strong genetic component underlying ASD has been firmly established from various lines of studies [4–7]. The search for ‘causative’ gene(s) has resulted in >10 whole genome scans reporting numerous putative linkage regions for ASD susceptibility [8–9]. Genetic association studies have identified numerous candidate genes for ASD [10–12]; however, most candidates fail to replicate between studies and populations. In a minor proportion of cases, chromosomal aberrations have been identified [13]. Recently, submicroscopic copy number variations (CNVs) were strongly associated with ASD [8,14–15]. Additionally, ASD is consistently associated with a number of specific genetic disorders such as Fragile X Syndrome [16–17]. Single gene mutations are also linked to rare cases of ASD [18–19].

Together, hundreds of diverse genetic loci gathered from high throughput studies have been implicated in this disorder. Addressing the complexity of ASD, we have developed AutDB [20–21], a publicly available web-portal for on-going collection, manual curation, and visualization of genes linked to the disorder. First released by our laboratory in 2007, AutDB is widely used by both individual laboratories [22–25] and consortia (Simons Foundation) [26] for understanding genetic bases of ASD.

Functional studies for isolated candidate genes have provided important insight into ASD but are largely restricted to rare monogenic forms of the disorder [27–28]. Here, we have exploited the genetic heterogeneity of ASD to create a predictive gene map for novel ASD candidate genes. To build this predictive map, we assembled a reference dataset of 84 ASD candidate genes from AutDB for dual profiling with both functional and expression analysis. We then used this dual profile to construct a predictive gene map for ASD which can be utilized in future research regarding pathogenesis of this complex psychiatric disorder.
Results

**AutRef84 as an ASD Gene Reference Dataset**

A reference dataset of ASD candidate genes was initially extracted from the autism gene database AutDB [20–21]. This resource provides systematic collection of candidate genes linked to ASD encompassing four genetic classifications 1) **Rare**: rare single gene variants, disruptions/mutations, and submicroscopic deletions/duplications directly linked to ASD, 2) ** Syndromic**: genes implicated in syndromes in which a significant subpopulation develops autistic symptoms, 3) **Association**: small risk-conferring candidate genes with common polymorphisms identified from genetic association studies in idiopathic ASD, and 4) **Functional**: functional candidates. Genes belonging to more than one category are classified with both names.

In order to generate a high-confidence predictive gene map for ASD, we restricted the ASD reference gene set to higher risk-conferring, more penetrant ASD candidate genes. We filtered out lower risk-conferring ASD candidate genes, including **Functional** candidates devoid of any experimentally determined genetic link with ASD, as well as **Association** genes, which have suggestive evidence linking them to ASD [25,29–30]. For instance, none of

**Figure 1. Integrated analysis of reference sets of genes Linked to ASD.** A reference dataset of ASD-linked genes (AutRef84) was assembled from the **Rare** and **Syndromic** categories of AutDB (http://www.mindspec.org/autdb.html), a publicly available portal for ongoing collection of genes linked to ASD. A) Distribution of genetic categories in AutDB. B) Reference sets were analyzed using structured biological knowledge provided by Gene Ontology (GO) consortium [31].

doi:10.1371/journal.pone.0028431.g001
the genes in AutDB belonging solely to the Association category have reached genome-wide GWAS with independent replication or meta-analysis. The vast majority of candidate genes identified from genetic association studies are unreplicated or underpowered.

The resulting ASD reference gene set, AutRef84, included 64 Rare and 20 Syndromic genes (Figure 1A), the total number of genes identified within these categories as of the data-freeze. The list of AutRef84 genes is presented in Table 1, whereas a more annotated version is provided as Table S1. The AutRef84 dataset encompasses well-studied candidates such as neuregulins (NGLN1, NGLN3, and NGLN4X), MECP2, FMR1, and TSC1/2, together with lesser-known genes with single reports including RPL10, CACNA1G, and DPP6. Hence, AutRef84 captures the broad landscape of ASD-linked genes suitable for applying statistical analysis to derive common gene functions.

We then screened the human genome with both branches of this AutRef84 consisting of both functional and expression analyses.

Table 1. Functional Profile: Common Biological Functions in AutRef84

<table>
<thead>
<tr>
<th>Rare</th>
<th>Syndromic</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANKR11</td>
<td>CNTR4</td>
</tr>
<tr>
<td>AGAP1</td>
<td>CNTRAP2</td>
</tr>
<tr>
<td>APC</td>
<td>DLGAP2</td>
</tr>
<tr>
<td>ASTN2</td>
<td>DPP10</td>
</tr>
<tr>
<td>AUTS2</td>
<td>DPP6</td>
</tr>
<tr>
<td>BZRAP1</td>
<td>DPPD</td>
</tr>
<tr>
<td>C3orf58</td>
<td>E4F4E</td>
</tr>
<tr>
<td>CA6</td>
<td>FAPB5</td>
</tr>
<tr>
<td>CACNA1H</td>
<td>FAPB7</td>
</tr>
<tr>
<td>CADM1</td>
<td>FBXO40</td>
</tr>
</tbody>
</table>

We then applied AutRef84 to generate a predictive gene map from genetic association studies are unreplicated or underpowered. The vast majority of candidate genes identified from genetic association studies are unreplicated or underpowered. The list of AutRef84 genes is presented in Table 1, whereas a more annotated version is provided as Table S1. The AutRef84 dataset encompasses well-studied candidates such as neuregulins (NGLN1, NGLN3, and NGLN4X), MECP2, FMR1, and TSC1/2, together with lesser-known genes with single reports including RPL10, CACNA1G, and DPP6. Hence, AutRef84 captures the broad landscape of ASD-linked genes suitable for applying statistical analysis to derive common gene functions.

We then screened the human genome with both branches of this AutRef84 consisting of both functional and expression analyses.

We further characterized functionality of the AutRef84 gene set by conducting pathway analysis with Pathway Express [33]. Using the KEGG database, we derived five significantly enriched molecular pathways: cell adhesion molecules (6 genes, P = 7.5 × 10−4), mTOR signaling pathway (4 genes, P = 3.3 × 10−4), calcium signaling pathway (5 genes, P = 4.4 × 10−4), p53 signaling pathway (3 genes, P = 1.8 × 10−5), and MAPK signaling pathway (5 genes, P = 2.6 × 10−5) (Table S3). One of these pathways, cell adhesion molecules, is shared exactly with the GO enrichment analysis. The other categories are shared indirectly: calcium signaling pathway is included in the GO enrichment category cation transport and cation channel activity, while the mTOR, p53, and MAPK signaling pathways relate to the GO enrichment category negative regulation of signaling transduction.

To visualize the relationship among enriched GO categories, we generated directed acyclic graphs based on GO knowledge structure. The terminal nodes of these GO Trees, devoid of any ancestors provide a framework for semantics of candidate gene functions in AutRef84. Within the BP graph (Figure 2), enriched terminal nodes relate to ion channel activity (sodium ion transport: 5 genes, P = 6.9 × 10−4; negative regulation of signal transduction: 6 genes, P = 3.3 × 10−4; regulation of system process: 6 genes, P = 6.0 × 10−4; regulation of phosphate metabolic process: 8 genes, P = 9.7 × 10−4), neurogenesis/projection (neuron projection development: 6 genes, P = 4.8 × 10−4), or cell adhesion (cell adhesion: 11 genes; P = 8.4 × 10−4). Notably, the largest structural component of the BP GO Tree is connected to neuron projection development, which includes the enriched GO categories of neuron differentiation (9 genes, P = 6.0 × 10−4), neurogenesis (10 genes, P = 4.0 × 10−4), and central nervous system development (9 genes, P = 4.0 × 10−4). Within the AutRef84 MF graph (Figure 1A), enriched terminal nodes also relate to ion channel activity (voltage-gated cation channel activity: 6 genes, P = 1.1 × 10−4; cation transport (BP): 10 genes, P = 7.9 × 10−4; neurogenesis (BP): 10 genes, P = 4.0 × 10−5; voltage-gated channel activity (MF): 6 genes; P = 3.4 × 10−4; and synapse (CC): 7 genes; P = 6.2 × 10−4).

Table 1. AutRef84: A Reference set of Rare and Syndromic ASD-linked genes.
Table 2. Functional Profile of AutRef84: Enriched Gene Ontology (GO) Categories.1

<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO Term</th>
<th>P- Value</th>
<th>Odds Ratio</th>
<th>Exp Count</th>
<th>AutRef84 Genes</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASD BP</td>
<td>GO:0007417 Central Nervous System Development</td>
<td>4.00E-05</td>
<td>6.235</td>
<td>1.656</td>
<td>9</td>
<td>303</td>
</tr>
<tr>
<td></td>
<td>GO:0022008 Neurogenesis</td>
<td>4.00E-05</td>
<td>5.485</td>
<td>2.104</td>
<td>10</td>
<td>385</td>
</tr>
<tr>
<td></td>
<td>GO:0030182 Neuron Differentiation</td>
<td>6.00E-05</td>
<td>5.847</td>
<td>1.760</td>
<td>9</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td>GO:0007155 Cell Adhesion</td>
<td>1.10E-04</td>
<td>3.907</td>
<td>3.902</td>
<td>13</td>
<td>714</td>
</tr>
<tr>
<td></td>
<td>GO:0009968 Negative Regulation of Signal Transduction</td>
<td>3.30E-04</td>
<td>7.180</td>
<td>0.929</td>
<td>6</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>GO:0031175 Neuron Projection Development</td>
<td>4.80E-04</td>
<td>6.646</td>
<td>1.000</td>
<td>6</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td>GO:0044057 Regulation of System Process</td>
<td>6.00E-04</td>
<td>6.355</td>
<td>1.044</td>
<td>6</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td>GO:0019220 Regulation of Phosphatase Metabolic Process</td>
<td>9.70E-04</td>
<td>4.344</td>
<td>2.049</td>
<td>8</td>
<td>375</td>
</tr>
<tr>
<td></td>
<td>GO:0006814 Sodium Ion Transport</td>
<td>6.90E-04</td>
<td>7.818</td>
<td>0.705</td>
<td>5</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>GO:0006812 Cation Transport</td>
<td>7.90E-04</td>
<td>3.740</td>
<td>3.012</td>
<td>10</td>
<td>536</td>
</tr>
<tr>
<td>MF</td>
<td>GO:0022843 Voltage-Gated Cation Channel Activity</td>
<td>7.00E-05</td>
<td>9.662</td>
<td>0.699</td>
<td>6</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>GO:0031402 Sodium Ion Binding</td>
<td>2.60E-04</td>
<td>9.760</td>
<td>0.570</td>
<td>5</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>GO:0022832 Voltage-Gated Channel Activity</td>
<td>3.40E-04</td>
<td>7.098</td>
<td>0.938</td>
<td>6</td>
<td>181</td>
</tr>
<tr>
<td>CC</td>
<td>GO:0034703 Cation Channel Complex</td>
<td>6.00E-05</td>
<td>9.980</td>
<td>0.675</td>
<td>6</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>GO:0045202 Synapase</td>
<td>6.20E-04</td>
<td>5.309</td>
<td>1.454</td>
<td>7</td>
<td>269</td>
</tr>
<tr>
<td>Diabetes BP</td>
<td>GO:0065008 Regulation of Biological Quality</td>
<td>2.49E-09</td>
<td>9.553</td>
<td>2.772</td>
<td>16</td>
<td>948</td>
</tr>
<tr>
<td></td>
<td>GO:0003013 Circulatory System Process</td>
<td>2.67E-09</td>
<td>17.334</td>
<td>0.747</td>
<td>10</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>GO:0009893 Positive Regulation of Metabolic Process</td>
<td>6.80E-08</td>
<td>7.801</td>
<td>2.450</td>
<td>14</td>
<td>662</td>
</tr>
<tr>
<td></td>
<td>GO:0016265 Death</td>
<td>8.95E-04</td>
<td>3.552</td>
<td>3.706</td>
<td>11</td>
<td>963</td>
</tr>
<tr>
<td></td>
<td>GO:0012501 Programmed Cell Death</td>
<td>1.55E-03</td>
<td>3.503</td>
<td>3.356</td>
<td>10</td>
<td>872</td>
</tr>
<tr>
<td>MF</td>
<td>GO:0005179 Hormone Activity</td>
<td>6.26E-03</td>
<td>8.644</td>
<td>0.376</td>
<td>3</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>GO:0019900 Kinase Binding</td>
<td>9.82E-03</td>
<td>7.293</td>
<td>0.443</td>
<td>3</td>
<td>119</td>
</tr>
<tr>
<td>CC</td>
<td>GO:0000267 Cell Fraction</td>
<td>2.00E-03</td>
<td>3.608</td>
<td>2.881</td>
<td>9</td>
<td>805</td>
</tr>
<tr>
<td></td>
<td>GO:0031090 Organelle Membrane</td>
<td>4.62E-03</td>
<td>3.419</td>
<td>2.671</td>
<td>8</td>
<td>787</td>
</tr>
<tr>
<td>Control* BP</td>
<td>GO:0032940 Secretion By Cell</td>
<td>2.36E-05</td>
<td>4.921</td>
<td>2.856</td>
<td>12</td>
<td>524</td>
</tr>
<tr>
<td></td>
<td>GO:0010647 Positive Regulation Of Cell Communication</td>
<td>4.50E-05</td>
<td>4.964</td>
<td>2.570</td>
<td>11</td>
<td>478</td>
</tr>
<tr>
<td></td>
<td>GO:0023056 Positive Regulation Of Signaling</td>
<td>4.76E-05</td>
<td>4.931</td>
<td>2.586</td>
<td>11</td>
<td>481</td>
</tr>
<tr>
<td></td>
<td>GO:0010740 Positive Regulation Of Intracellular Protein Kinase Cascade</td>
<td>6.29E-05</td>
<td>6.444</td>
<td>1.365</td>
<td>8</td>
<td>254</td>
</tr>
<tr>
<td></td>
<td>GO:0018130 Heterocycle Biosynthetic Process</td>
<td>8.27E-06</td>
<td>6.727</td>
<td>1.741</td>
<td>10</td>
<td>343</td>
</tr>
<tr>
<td></td>
<td>GO:0034654 Nucleobase, Nucleoside, Nucleotide And Nucleic Acid Biosynthetic Process</td>
<td>1.84E-05</td>
<td>6.881</td>
<td>1.513</td>
<td>9</td>
<td>298</td>
</tr>
<tr>
<td></td>
<td>GO:004283 Small Molecule Biosynthetic Process</td>
<td>6.05E-05</td>
<td>4.467</td>
<td>3.158</td>
<td>12</td>
<td>622</td>
</tr>
<tr>
<td></td>
<td>GO:0009165 Nucleotide Biosynthetic Process</td>
<td>7.63E-05</td>
<td>6.472</td>
<td>1.406</td>
<td>8</td>
<td>277</td>
</tr>
<tr>
<td></td>
<td>GO:0051090 Regulation Of Transcription Factor Activity</td>
<td>2.94E-05</td>
<td>7.432</td>
<td>1.226</td>
<td>8</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>GO:0051098 Regulation Of Binding</td>
<td>3.30E-05</td>
<td>6.341</td>
<td>1.624</td>
<td>9</td>
<td>298</td>
</tr>
<tr>
<td></td>
<td>GO:0043254 Regulation Of Protein Complex Assembly</td>
<td>2.69E-05</td>
<td>11.554</td>
<td>0.591</td>
<td>6</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>GO:0031329 Regulation Of Cellular Catabolic Process</td>
<td>7.47E-05</td>
<td>6.488</td>
<td>1.401</td>
<td>8</td>
<td>272</td>
</tr>
</tbody>
</table>
P = 7.0 × 10^{-5}; sodium ion binding: 5 genes, P = 2.6 × 10^{-4}). Similarly, within the AutRef84 CC graph (Figure S2), enriched terminal nodes describe cellular components important for ion channel activity (cation channel complex: 6 genes, P = 6.0 × 10^{-4}) or ion channel activity/cell adhesion (synapse: 7 genes, P = 6.2 × 10^{-4}).

Expression Profile: Region-Specific Enrichment of AutRef84

We next performed tissue-specific expression analysis of AutRef84 using the DAVID bioinformatics resource. We discovered nine anatomical regions of highly enriched ASD candidate gene expression (P ≤ 1.0 × 10^{-4}; accounting for multiple testing), including four brain regions: 1) the olfactory bulb (33 genes, P = 1.7 × 10^{-6}), which transmits smell; 2) the occipital lobe (32 genes, P = 1.0 × 10^{-6}), which processes visual information; 3) the prefrontal cortex (25 genes, P = 3.1 × 10^{-6}), which is important for executive function; and 4) the pituitary (26 genes, P = 3.2 × 10^{-5}), which regulates hormone secretion (Figure 3). The enrichment of ASD-linked genes within non-brain regions is not surprising, given the pleiotropic expression of genes within the human body. However, due to the categorization of ASD as a neuropsychiatric illness, we focused exclusively on enriched brain regions. A list of AutRef84 genes expressed within each region is provided in Table S4.

We then performed network representation to illustrate relationships of AutRef84 gene expression among these shared brain regions (Figure 4). Of the AutRef84 gene set, a total of 45 genes were enriched among these four brain regions. A core set of 16 genes showed enriched expression in all four brain regions: A2BP1, APC, CNTNAP2, DPP6, FABP7, IL1RAPL1, NBEA, NF1, NLGN3, NLGN4X, NRXN1, PCDH9, RAPGEF4, RIMS3, SCN2A, and SEZ6L2. Additionally, eight genes were expressed in three regions (including the Rett Syndrome gene MECP2 and the Fragile X Syndrome gene FMR1), ten genes were expressed in two regions, and eight genes were expressed in only one of these enriched regions.

Disease Specificity of AutRef84 Dual Profile

To examine disease specificity of the AutRef84 dual profile, we compared it to both 1) a diabetes set of 54 genes serving as a disease-specific control (Table S5), 2) a non-specific disease reference set of 78 genes generated by randomly sampling genes the OMIM database which did not show significant association to any one particular disease, and 3) 1000 control gene sets of size n = 84 that were randomly sampled from the OMIM database. First, we analyzed AutRef84 functional profile specificity. By performing GO enrichment analysis with the same filtering criteria as above, we found that the diabetes gene set was enriched within nine GO categories (five BP categories, two MF categories, and two CC categories) (Table 2), and the 1000 random control gene sets were enriched within 18 GO categories (all BP categories) (Table 2). None of the top enriched GO terms were shared among ASD, diabetes, and control datasets. Whereas the top enriched GO terms for ASD were concentrated in synaptic functions, the diabetes reference set was enriched for metabolic cellular processes, and the control reference sets were distributed primarily across various signaling pathways.

Second, we examined AutRef84 expression profile specificity. For this analysis, we used compared AutRef84 with the diabetes gene set as well as a non-specific disease gene set assembled by randomly sampling OMIM (Table S6). Region-specific enrichment of the diabetes gene set was restricted to the pons, whereas expression of the non-specific disease dataset was concentrated in the uterus and three brain regions distinct from that of ASD (Figure 3). Like the functional profile, no enrichment was shared among ASD, diabetes, and non-specific disease datasets. Although each gene set showed enrichment in distinct brain regions, it is noteworthy that statistical significance of the enriched ASD brain regions was generally two orders of magnitude higher than those of diabetes and non-specific disease datasets. Eight of the nine ASD regions satisfy a more stringent P-value of 0.001 (Figure 3, dotted line), whereas none of the diabetes and control regions meet this cut-off, implying that brain region enrichment of ASD genes is much less likely to occur by chance.

Predictive Gene Map Generated with Dual AutRef84 Profile

The creation of AutRef84 provides a usable framework for systems biology analysis of molecular events underlying ASD pathogenesis. Here we applied the dual AutRef84 profile to generate a predictive gene map for ASD. First we used the functional AutRef84 profile to screen the human genomic sequence at Ensembl database (NCBI build 36) with the biomaRt package of Bioconductor. We identified 1185 genes matching the functional AutRef84 profile (Table S7), and their genome-wide distribution pattern is illustrated in Figure 5. This map indicates

<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO Term</th>
<th>P-Value</th>
<th>Odds Ratio</th>
<th>Exp Count</th>
<th>AutRef84 Genes</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0009141</td>
<td>Nucleoside Triphosphate Metabolic Process</td>
<td>8.57E-06</td>
<td>6.020</td>
<td>2.153</td>
<td>11</td>
<td>418</td>
</tr>
<tr>
<td>GO:0009205</td>
<td>Purine Ribonucleoside Triphosphate Metabolic Process</td>
<td>4.20E-05</td>
<td>5.505</td>
<td>2.102</td>
<td>10</td>
<td>408</td>
</tr>
<tr>
<td>GO:0009259</td>
<td>Ribonucleotide Metabolic Process</td>
<td>8.22E-05</td>
<td>5.058</td>
<td>2.277</td>
<td>10</td>
<td>442</td>
</tr>
<tr>
<td>GO:0071844</td>
<td>Cellular Component Assembly At Cellular Level</td>
<td>5.31E-05</td>
<td>4.004</td>
<td>4.165</td>
<td>14</td>
<td>797</td>
</tr>
<tr>
<td>GO:0051336</td>
<td>Regulation Of Hydrolase Activity</td>
<td>6.90E-05</td>
<td>4.716</td>
<td>2.692</td>
<td>11</td>
<td>494</td>
</tr>
<tr>
<td>GO:0006259</td>
<td>DNA Metabolic Process</td>
<td>1.42E-05</td>
<td>4.856</td>
<td>3.199</td>
<td>13</td>
<td>621</td>
</tr>
</tbody>
</table>

BP = Biological Process, MF = Molecular Function, CC = Cellular Component.
* = 1000 gene sets of n = 84, randomly assembled from OMIM.

doi:10.1371/journal.pone.0028431.t002
uneven distribution, with dense packing of matched genes in 13 discrete chromosomal regions that reached statistical significance using a DAVID analysis with a $P$-value cut-off value of $p<0.05$ (Table S8).

We then applied the AutRef84 expression profile to filter this initial predictive gene set. Using network representation of shared brain expression within the set of 1185 genes described above, we defined a subset of 460 genes matching both functional and expression profiles of AutRef84 (Figure 6). Of this subset, 159 genes are expressed in all four enriched brain regions, 62 genes were common to three of the enriched brain regions, 89 genes overlapped in two of the enriched regions, and 150 genes were expressed in only one of the enriched brain regions (Table S9).

Importantly, the accuracy of the final predictive gene map was demonstrated by correctly capturing 18 existing ASD-associated genes which were not part of the AutRef84 input dataset (Table S10): 13 genes linked to ASD by genetic association studies (therefore part of the Association category of AutDB), three genes whose function is relevant to ASD (therefore included in the Functional category of AutDB), and two Rare/Syndromic genes that have been discovered since the original AutRef84 data-freeze. Some of these candidate genes are particularly interesting candidates for ASD, such as GABRB3, a GABA receptor subunit linked to ASD by multiple association studies [34–41], NPAS2, a transcription factor involved in circadian rhythms that has been associated with ASD [42]; RELN, an extracellular matrix protein involved in cell migration whose association with ASD has been replicated [43–46]; and SEMA5A, an axon guidance molecule shown to be downregulated in ASD [47]. However, the remaining 442 genes have not previously been associated with ASD and form a novel pool of potential ASD candidate genes.

**Discussion**

In this report, we have defined the first composite reference profile of ASD candidate genes, AutRef84. In contrast to another
recent ASD profile based on functional annotation of candidate genes [48], here we created a dual reference profile of AutRef84 by performing both functional and expression analyses of ASD candidate genes. Derived from data extracted from >158 references, AutRef84 consolidates knowledge about Rare and Syndromic ASD genes, whose relationship to ASD has been firmly established. For the functional profile, we conducted GO enrichment to discover that AutRef84 genes are enriched in biological functions related to three major areas: 1) neurogenesis/projection, 2) cell adhesion, and 3) ion channel activity. For the expression profile, we analyzed tissue-specific expression patterns to find that AutRef84 genes are enriched in brain regions vital to sensory information processing (olfactory bulb, occipital lobe), executive function (prefrontal cortex), and hormone secretion (pituitary). We then applied this dual profile to create a genome-wide predictive gene map for ASD consisting of 460 putative candidate genes. Of these 460 genes, 18 were previously associated with ASD but were included in our input AutRef84 dataset, demonstrating the predictive power of this gene map. The remaining 442 genes are entirely novel putative ASD risk genes. Together, our predictive gene map can serve as a tool for researchers to prioritize molecular pathways underlying ASD pathogenesis, thereby accelerating the discovery of targeted treatments for this disorder.

Figure 3. AutRef84 expression profile: region-specific enrichment of gene expression. Analysis of tissue expression profiles for AutRef84 genes using the DAVID bioinformatics tool (http://david.abcc.ncifcrf.gov/) demonstrates region-specific enrichment with high statistical significance (p<0.0001) in four areas of the central nervous system: olfactory bulb, occipital lobe, prefrontal cortex, and pituitary. Whereas the olfactory bulb and occipital lobe are involved in sensory processing (smell and vision, respectively), the prefrontal cortex controls executive function and the pituitary gland directs hormone secretion. None of the enriched regions overlapped with those of diabetes or non-specific disease gene sets. * = Bonferroni corrected.
doi:10.1371/journal.pone.0028431.g003

Our functional profile revealed that ASD candidate genes are concentrated in three biological processes critical for synaptic transmission: neurogenesis/projection, cell adhesion, and ion channel activity. A ‘synaptic dysfunction’ hypothesis for ASD is widely acknowledged [49–50]. However, molecular support for this hypothesis rests mainly on cell adhesion binding partners neuroligins (NLGN3, NLGN4X) and neurexins (NRXN1), as well as the scaffolding protein SHANK3 – all identified in rare cases of ASD. The availability of curated, annotated datasets of ASD-linked genes provides unique computational opportunities to identify common biological functions associated with these genes [25]. Here, we use a reference set of genes, rigorous statistical analysis, and comparative analysis with multiple control datasets to provide molecular support for synaptic bases of ASD.

For instance, the largest concentration of synaptic categories for ASD-linked genes involves ion regulation. Six of the 15 enriched GO categories for AutRef84 were sodium transport, cation transport, voltage-gated cation channel, sodium ion binding, voltage-gated channel activity, and cation channel complex. This unbiased study of ASD candidate genes supports a previously established theory of ASD pathogenesis proposing an increased excitation:inhibition ratio [51]. In correspondence with this theory, approximately 10–30% of individuals with autism are also diagnosed with epilepsy [52–53], a disease caused by ion channel dysfunction. To further examine the role of ion channels in both diseases, future studies should compare the functional profile of AutRef84 with one created from an epilepsy reference gene set.

Another major component of our ASD reference profile is neurogenesis/projection. Enriched GO categories within this neurobiological classification included neurogenesis, neuron differentiation, neuron projection development, central nervous system development, and cell adhesion. Impairments of these neurodevelopmental processes may contribute to accelerated head growth observed in children with ASD [54–55]. Additionally, neurogenesis continues to play a role in adult function of brain regions such as the hippocampus [56] and amygdala [57]. Inability of neurons to regenerate within
these brain areas may lead to deficits in emotional processing observed in autism [58–59].

Our expression profile defines four critical brain regions of ASD pathogenesis: olfactory bulb, occipital lobe, prefrontal cortex, and pituitary. Dysfunction of each of these brain areas in ASD has been suggested by previous functional evidence. For example, the olfactory bulb, which transmits information pertaining to smell, has been strongly implicated in mouse models of ASD due to its well-established role in their social behavior [60]. Interestingly, humans with ASD also exhibit altered olfactory perception [61–62]. Electrical abnormalities have been observed in the occipital lobe of ASD individuals [63], suggesting that impaired facial recognition associated with ASD [64] may at least partially be due to altered visual processing. The prefrontal cortex is critical for executive function skills deficient in ASD, such as decision-making, attention, and working memory. In support, ASD individuals exhibit decreased activation of the prefrontal cortex when performing cognitive tasks [65]. Finally, impaired hormone secretion by the pituitary has long been proposed to contribute to ASD [66], although recent studies have highlighted the potential importance of hormones underlying social behavior, such as oxytocin and vasopressin [67].

Although our AutReF84 expression profile highlights anatomical regions likely to be involved in ASD pathogenesis, it should be interpreted with caution. The four brain regions described above (olfactory bulb, occipital lobe, prefrontal cortex, and pituitary) are the only ones which survived multiple testing in our statistical analysis, but previous evidence suggests that other brain regions functionally relevant to ASD such as the amygdala or cerebellum may also be involved [58,68]. Likewise, some AutReF84 genes were enriched in non-brain regions, reflecting the pleiotropic expression of genes within the human body. Notably, although expression profiles of diabetes and control datasets also showed enrichment in some brain regions, these brain regions were distinct from of ASD genes and, more importantly were two orders of magnitude less statistically significant. Together, disease specificity of the AutReF84 dual profile indicates the utility of disease-based reference profiling.

Notably, the results of our computational analysis match evidence generated by single gene studies. For example, our expression profile identified the cell adhesion molecule CNTNAP2 as one of the core set of 16 AutReF84 genes enriched in all four significant brain regions, prioritizing it as a high-confidence ASD candidate gene. In support, one recent neuroimaging study used magnetic resonance imaging (MRI) and diffusion tensor imaging to demonstrate that subjects with an ASD-associated single nucleotide polymorphism in CNTNAP2 showed a significant reduction in grey and white matter volume of the occipital and frontal lobes compared with controls [69]. Likewise, a newly published functional MRI showed that another ASD-associated single nucleotide polymorphism of CNTNAP2 altered functional connectivity within the frontal lobe [70]. Additional functional studies will be critical for defining the contribution of our prioritized gene set to molecular pathways dysfunctional in ASD.

In conclusion, our predictive gene map for ASD is a valuable tool by which to prioritize the field of ASD genomics. Our composite reference profile of AutReF84 also provides insight into the molecular etiology of autism, with important implications for drug development. Moreover, our construction and evaluation of AutReF84 can act as a general model for consolidating collective knowledge of a complex disorder into a usable framework of common biological functions.

Figure 4. Network analysis of the AutReF84 expression profile. In this visual representation of the network, each group of gene nodes is spatially positioned near the brain region or regions in which the genes are expressed. The color of each group of gene nodes was derived by averaging red, green, and blue values of the colors of the linked brain region nodes. doi:10.1371/journal.pone.0028431.g004
Materials and Methods

Compilation of Gene ASD and Control Reference Sets

We have developed an autism gene database, AutDB [71,20–21], for ongoing cataloguing of genes linked to ASD. A comprehensive collection of ASD-linked genes was initially compiled from an exhaustive search of the scientific literature from PubMed database at NCBI [72]. The search terms included ‘gene’ AND (‘autism’ OR ‘autistic’) restricted to the titles and abstracts of the publication for retrieval. Furthermore, candidate genes listed in review articles on molecular genetics of ASD, along with cross-references therein, were mapped and added (if new) to our candidate gene list from PubMed searches to compile the most exhaustive gene set. After its first release (Jan 1, 2007), a daily semi-automated search of PubMed with the same keywords was performed to maintain an up-to-date resource of all candidate genes linked to ASD. Additionally, relevant journal articles in the fields of genetics, neurobiology, and psychiatry were screened on a regular basis to enrich the resource. AutRef84 assembled with a data-freeze of May 2010. The authors individually verified all candidate genes included in the reference dataset by reading the full-text primary reference article linking the candidate gene to ASD.

Non-ASD gene sets were compiled using the Online Mendelian Inheritance in Man (OMIM) database [73]. The diabetes dataset consisted of 54 genes verified for linkage association with diabetes and expression in Beta Cells/Islets in the Type 1 Diabetes Database [74] and manually analyzed to exclude any genes whose link to Diabetes was based on genetic association studies.

The non-specific disease reference set was curated by generating a random sampling of 78 genes from the OMIM database which did not show significant association to any one particular disease. The 1000 control gene sets of n = 84 were assembled by randomly sampling the OMIM database.

Bioconductor Analysis

Enrichment of GO categories was performed using the Conditional HyperGTest in the annotation background of hgu133a as described in the GOStats vignette (S.Falcon and R. Gentleman, October 3, 2007). The Conditional HyperG Test uses the structure of the GO graph to estimate for each term whether or not there is evidence beyond that which is provided by the term’s children to designate the term statistically over-represented. The algorithm conditions on all child terms also significant at the specified $P$-value cut-off. Given a subgraph of one of the three GO ontologies, the terms with no child categories are tested first, followed by the nodes whose children have already been tested. If any of a given node’s children tested significant, the appropriate conditioning is performed.

Results of the Conditional HyperG Test were analyzed and visualized in an Excel spreadsheet for GO category, $P$-value, Odds ratio, Expected count, AutRef84 gene count, and annotation category size (Table S8). The hierarchical relationship between enriched GO terms was visualized by constructing directed acyclic graphs using GOStats package in Bioconductor (Figure 3; Figures S1 and S2). Terminal leaves of the graphs were extracted for analysis. The complete list of packages used for GO analysis is shown in the Methods S1 section.

DAVID Analysis

We used the Database for Annotation, Visualization, and Integrated Discovery (DAVID) version 6.7 [75] to identify annotation terms significantly enriched in each reference gene set. We used the modified Fisher’s exact test, or EASE score, to identify enriched annotation terms derived from GNF_U133A_QUARTILE and gene ontology (GO) annotation terms, which includes Biological Process (BP), Molecular function (MF), and Cellular Component (CC) categories. We used the more specific GO term categories provided by DAVID, called GO FAT, to minimize the redundancy of the more general GO terms in the analysis to increase the specificity of the terms.

A list of gene symbols was generated for each dataset and used as input into DAVID. We used the Functional Annotation Tool, with the Human Genome U133A Plus 2.0 Array as the gene background, to independently analyze each gene set. We used a count threshold of 5 and the default value of 0.1 for the EASE score settings. We also used the Benjamini corrected $P$-value, with $p<0.05$ as the significance threshold. Significant annotation terms identified in the GNF annotation category were further filtered using the interquartile range of the category size, where the 1st and 3rd quartile were removed from the results. Significant annotation terms in the remaining GO annotation categories were filtered by removing those terms with a category size less than 100 and greater than 1000.

Genome-Wide Expression Profile

We used the biomaRt package of Bioconductor to screen the human genomic sequence at Ensembl database (NCBI build 36) with the optimized AutRef84 profile. For this analysis, hgu133a was used as the universe.

Network Visualization

To convey overlapping gene expression between these four regions, we produced a bipartite network consisting of AutRef84
ASD candidate genes and the four brain regions. We assigned links between the genes and their corresponding brain regions. We then assigned a category to each gene with respect to its linked brain regions. (For example, genes expressed in the occipital lobe and in the pituitary were placed in one category, while genes expressed only in the prefrontal cortex were placed in another, and so on.) Next, we used the attribute circle layout in Cytoscape [76] to arrange the nodes in each category into circles. Each circle was then manually repositioned in a location close to its linked brain region or regions. The four brain region nodes were assigned colors based on their positions in an RGB (red, green, blue) cube color space [77]. The color of the nodes in each gene category circle was derived by averaging R, G, and B values of the colors of the linked brain region nodes.

Supporting Information

Figure S1 AutRef84 functional profile: graphical representation of over-represented Molecular Function (MF) categories. Using Bioconductor, we generated directed acyclic graphs based on GO knowledge structure. Enriched GO categories of AutRef84 are represented by rectangular boxes. Terminal nodes are illustrated in yellow. Similar to the AutRef84 BP GO Tree (Figure 2), enriched terminal nodes also relate to ion channel activity (voltage-gated cation channel activity: 6 genes, \( P = 7.0 \times 10^{-3} \); sodium ion binding: 5 genes, \( P = 2.6 \times 10^{-4} \)).

(PDF)

Table S1 Expanded details of AutRef84 gene set.
(PDF)

Table S2 Enriched GO categories of AutRef84 using DAVID analysis.
(PDF)

Table S3 KEGG pathway analysis of AutRef84 using Onto Express.
(PDF)
Table S4 List of AutRef84 genes expressed within enriched regions.

Table S5 Reference set of diabetes-linked genes.

Table S6 Non-specific disease gene set.

Table S7 Set of 1185 predicted ASD candidate genes matching the AutRef84 functional profile.

Table S8 Significantly enriched cytotag categories for the predictive set of 1185 genes using DAVID analysis.

Table S9 Set of 460 predicted ASD candidate genes matching the AutRef84 dual profile.

Table S10 Previously identified ASD-linked genes matching the AutRef84 dual profile which were not included in the input dataset.

**References**


