1 Introduction

This is a release PDF for denovo-db version 1.5. This document describes studies by phenotype, lists studies previously included but now removed, discusses studies considered but not included, and provides an appendix describing where the variants were pulled from each paper. Please see our Data Usage and Release page at http://denovo-db.gs.washington.edu/denovo-db/Usage.jsp. We invite you to email our mailing list at denovo-db@uw.edu with any comments or questions.

2 Description of studies

2.1 Studies by phenotype

In the following subsections we will describe the studies in denovo-db 1.5 in alphabetical order based on their primary phenotype.

2.1.1 Acromelic frontonasal dysostosis

Smith et al. 2014 is the only study in the database on acromelic frontonasal dysostosis. It contains data from four probands of which three were whole-exome sequenced.

2.1.2 Amyotrophic lateral sclerosis

There are two studies (Table 1) in denovo-db on amyotrophic lateral sclerosis (ALS). As far as we know, there are no overlaps between the two cohorts.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steinberg2015</td>
<td>44</td>
</tr>
<tr>
<td>Chesi2015</td>
<td>47</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>91</strong></td>
</tr>
</tbody>
</table>

2.1.3 Anophthalmia/Microphthalmia

There is only one study (Slavotinek2015) on anophthalmia/microphthalmia in denovo-db and it contains 25 probands.
2.1.4 Autism

There are a number of autism whole-exome, whole-genome, and targeted sequencing studies. Table 2 contains a list of all of these studies. The data come from different cohorts studied by many researchers. These include the Simons Simplex Collection (SSC), The Autism Simplex Collection (TASC), the Autism Sequencing Consortium (ASC), MSSNG, and the National Institute of Mental Health (NIMH) as well as clinical cohorts. For aggregate studies of the autism data it is important to know the sample overlaps and what to consider for analysis (Table 3).

First, we highly recommend that the targeted sequencing studies (ASD1,2, ASD3) be ignored except for individuals from the SSC (include .p1 in their name) who are present there because of their validation status. denovo-db picks variant representation in studies based on validation status so SSC variants were detected in exome but validated by targeted sequencing studies.

Second, the Michaelson et al. 2012 study is a set of 10 monozygotic twin pairs and for aggregate statistics only one set of the variants should be represented in analysis.

Third, although the overall denominator reported is 2,270 in de Rubeis et al. 2014 there were 825 SSC probands included so we report the difference of 1,445 in denovo-db since the SSC proband results represented in that study are present in other studies in the database.

Fourth, the Yuen2016 and Yuen2017 studies are both in the MSSNG dataset and therefore the denominator should be adjusted to Yuen2017’s 1,625 total number. Also, the Yuen studies contain related affected children which should not matter for de novo variants except for the less common case of germline mosaicism.

In sum there are a total of 5,624 unique individuals (Table 3) with autism from exome and genome sequencing studies.

Table 2: Autism studies in denovo-db 1.5

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Name</th>
<th>PMID</th>
<th>Probands</th>
<th>Controls</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>autism</td>
<td>Iossifov</td>
<td>25363768</td>
<td>2508</td>
<td>1911</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Krumm</td>
<td>25961944</td>
<td>2377</td>
<td>1786</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>ASD1,2</td>
<td>23160955</td>
<td>2446</td>
<td>0</td>
<td>targeted</td>
</tr>
<tr>
<td></td>
<td>ASD3</td>
<td>25418537</td>
<td>3486</td>
<td>2493</td>
<td>targeted</td>
</tr>
<tr>
<td></td>
<td>Turner2016</td>
<td>26749308</td>
<td>53</td>
<td>40</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>DeRubeis2014</td>
<td>25363760</td>
<td>2270</td>
<td>0</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Michaelson2012</td>
<td>23260136</td>
<td>10</td>
<td>0</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Lee2014</td>
<td>24501278</td>
<td>1</td>
<td>0</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Tavassoli2014</td>
<td>24650168</td>
<td>1</td>
<td>0</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Yuen2016</td>
<td>27525107</td>
<td>200</td>
<td>0</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Yuen2017</td>
<td>28263302</td>
<td>1625</td>
<td>2</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Hashimoto2015</td>
<td>26582266</td>
<td>30</td>
<td>0</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Moreno-Ramos2015</td>
<td>26352270</td>
<td>4</td>
<td>0</td>
<td>exome</td>
</tr>
</tbody>
</table>

2.1.5 Bipolar disorder

There is only one study on bipolar disorder (Kataoka2016) in denovo-db and it contains 79 probands.

2.1.6 Cantú syndrome

There is only one study (vanBon2012) on Cantú syndrome in denovo-db and it contains four probands with one that was exome sequenced.
Table 3: Unique exome/genome denominators for performing aggregate autism analysis

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Studies</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC</td>
<td>Iossifov, Krumm, Turner2016 ASD1/2/ASD3 (.s1 only)</td>
<td>2508</td>
</tr>
<tr>
<td>ASC</td>
<td>DeRubeis2014</td>
<td>1445</td>
</tr>
<tr>
<td>MSSNG</td>
<td>Yuen2016, Yuen2017</td>
<td>1625</td>
</tr>
<tr>
<td>NIMH</td>
<td>Michaelson2012</td>
<td>10</td>
</tr>
<tr>
<td>Clinical</td>
<td>Lee2014</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Tavassoli2014</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>Hashimoto2015</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Moreno-Ramos2015</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>5624</td>
</tr>
</tbody>
</table>

2.1.7 Congenital diaphragmatic hernia

There is only one study (Yu2015) on congenital diaphragmatic hernia in denovo-db and it contains 39 probands.

2.1.8 Congenital heart disease

There are three studies (Lifton, Homsy2015, Sifrim2016) on congenital heart disease in denovo-db. There are overlaps between these studies. By removing all sample overlaps that we could identify, there were a total of 2,072 probands.

2.1.9 Controls

There are a few studies where there were either only controls sequenced or cases and controls. These studies are shown in Table 4. The data from these studies come from different cohorts studied by many researchers. These include the SSC, MSSNG, Genome of the Netherlands (GoNL), and others. For aggregate studies of the control data, it is important to know the overlaps and what to consider for analysis (Table 5).

We highly recommend that the targeted sequencing study (ASD3) be ignored except for individuals from the SSC (include .s1 in their name) who are present there because of their validation status.

As an additional note, Conrad2011 and Ramu2013 both study only one sample (NA12878).

In sum up there are a total of 2,278 unique control individuals (Table 5) from exome and genome sequencing studies.

Table 4: Controls in denovo-db 1.5

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Name</th>
<th>PMID</th>
<th>Cases</th>
<th>Controls</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>Iossifov</td>
<td>25363768</td>
<td>2508</td>
<td>1911</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Krumm</td>
<td>25961944</td>
<td>2377</td>
<td>1786</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>ASD3</td>
<td>25418537</td>
<td>3486</td>
<td>2493</td>
<td>targeted</td>
</tr>
<tr>
<td></td>
<td>Turner2016</td>
<td>26749308</td>
<td>53</td>
<td>40</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Yuen2017</td>
<td>28263302</td>
<td>1625</td>
<td>2</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>GoNL</td>
<td>24974849</td>
<td>0</td>
<td>250</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Besenbacher2014</td>
<td>25597990</td>
<td>0</td>
<td>10</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Conrad2011</td>
<td>21666693</td>
<td>0</td>
<td>1</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Ramu2013</td>
<td>23975140</td>
<td>0</td>
<td>1</td>
<td>genome</td>
</tr>
<tr>
<td></td>
<td>Rauch2012</td>
<td>23020937</td>
<td>51</td>
<td>20</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Gulsuner2013</td>
<td>23911319</td>
<td>105</td>
<td>84</td>
<td>genome</td>
</tr>
</tbody>
</table>
Table 5: Unique exome/genome denominators for performing aggregate control analysis

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Studies</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC Iossifov, Krumm, Turner2016</td>
<td>ASD3 (.s1 only)</td>
<td>1911</td>
</tr>
<tr>
<td>MSSNG</td>
<td>Yuen2017</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>GoNL</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Besenbacher2014</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Conrad2011/Ramu2013</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Rauch2012</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Gulsuner2013</td>
<td>84</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2278</td>
</tr>
</tbody>
</table>

2.1.10 Developmental disorder

There is only one study on developmental disorder (DDD2017) in denovo-db and it contains 4,293 probands.

2.1.11 Early onset Alzheimer’s disease

There is only one study (Rovelet-Lecrux2015) on early onset Alzheimer’s disease in denovo-db and it contains 12 probands.

2.1.12 Early onset Parkinson’s disease

There is only one study (KunRodrigues2015) on early onset Parkinson’s disease in denovo-db and it contains 21 probands.

2.1.13 Epilepsy

There are five studies on epilepsy with a total of 532 probands. The studies are shown in Table 6.

Table 6: Unique denominators for performing aggregate epilepsy analysis

<table>
<thead>
<tr>
<th>Studies</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>epi4k2013</td>
<td>264</td>
</tr>
<tr>
<td>Veeramah2012</td>
<td>1</td>
</tr>
<tr>
<td>Barcia2012</td>
<td>3</td>
</tr>
<tr>
<td>Veeramah2013</td>
<td>10</td>
</tr>
<tr>
<td>Helbig2016</td>
<td>254</td>
</tr>
<tr>
<td>Total</td>
<td>532</td>
</tr>
</tbody>
</table>

2.1.14 Intellectual disability

There are four studies on intellectual disability with a total of 1,010 probands. The studies are shown in Table 7.

2.1.15 Neural tube defects

There is only one study (Lemay2015) on neural tube defects in denovo-db and it contains 43 probands.
Table 7: Unique denominators for performing aggregate intellectual disability analysis

<table>
<thead>
<tr>
<th>Studies</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>deLigt2012</td>
<td>100</td>
</tr>
<tr>
<td>Rauch</td>
<td>51</td>
</tr>
<tr>
<td>Halvardson2016</td>
<td>39</td>
</tr>
<tr>
<td>Lelieveld2016</td>
<td>820</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1010</td>
</tr>
</tbody>
</table>

2.1.16 Schizophrenia

There are five studies on schizophrenia with a total of 800 probands. The studies are shown in Table 8.

Table 8: Unique denominators for performing aggregate schizophrenia analysis

<table>
<thead>
<tr>
<th>Studies</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulser2013</td>
<td>105</td>
</tr>
<tr>
<td>McCarthy2014</td>
<td>57</td>
</tr>
<tr>
<td>Fromer2014</td>
<td>623</td>
</tr>
<tr>
<td>Smedemark-Margulies2016</td>
<td>1</td>
</tr>
<tr>
<td>Kranz2015</td>
<td>14</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>800</td>
</tr>
</tbody>
</table>

2.1.17 Sporadic infantile spasm syndrome

There is only one study (Dimassi2015) on sporadic infantile spasm syndrome in denovo-db and it contains 10 probands.

2.2 Studies previously in denovo-db and removed in version 1.5

2.2.1 Jiang et al. 2013, PMID: 23849776

Jiang et al. 2013 was a whole-genome sequencing study of 32 families with autism. We realized in the preparation of denovo-db 1.5 that there were 20 pairs of identical duplicates with Yuen et al. 2016. The Yuen paper did not mention previous sequencing of these families but the senior author is the same so we believe the samples overlap. The naming convention is similar between studies. For example, one of the overlaps is at chromosome 1, position 16376358, C to G change and is in sample 2-1342 proband in the Jiang study and sample 2-1342-003 in the Yuen study. To be conservative we removed the Jiang study from version 1.5.

2.2.2 Yuen et al. 2015, PMID: 25621899

Yuen et al. 2015 was removed from denovo-db 1.5 as it had considerable overlap with Yuen et al. 2017. The two studies have the same samples but used “.” in the sample names in one paper and “_” in the other paper therefore bypassing our original pipelines for catching duplicates. Also, some of the samples in the 2017 paper had an a or b at the end of their name, which may correspond to different sequencing platforms. We removed those from denovo-db 1.5.
2.3 Studies considered for denovo-db but not included in version 1.5

2.3.1 Neale et al. 2012, PMID: 22495311

We had hoped to add the Neale et al. 2012 study to the database, which studied the exomes of 175 families with autism. Unfortunately, we were not able to disentangle the sample overlaps between this paper and the de Rubeis 2014 paper published by some of the same authors. Because of potential overlap we excluded it from this version of the database.

2.3.2 DDD 2015, PMID:25533962

Because of potential overlap with DDD 2017 we excluded DDD 2015 from this version of the database.

3 Appendix

In the appendix we describe the location of the data in the original papers.

3.0.1 ASD1_2, PMID:23160955

The data from this paper comes from Tables S11 and S14.

3.0.2 ASD3, PMID:25418537

The data from this paper comes from Supplementary Data 1 and 2.

3.0.3 Barcia2012, PMID:23086397

The data from this paper comes from Supplementary Table 2 and the number of trios exome sequenced (n = 3) is described in the main text.

3.0.4 Besenbacher2014, PMID:25597990

The data from this paper comes from Supplementary Data 3. The count of n = 10 trios is from the abstract.

3.0.5 Chesi2015, PMID:23708140

The data from this study comes from Table 1 and the trio count of n = 47 is from the abstract.

3.0.6 Conrad2011, PMID:21666693

The data from this study comes from Table S1 and only the sites that were marked as germline de novo in sample NA12878 were retained for denovo-db. Since the coordinates were on hg18, we lifted them over to hg19.

3.0.7 DDD 2017, PMID:28135719

The data from this paper comes from Supplementary Table 1 and the count of families (n = 4,293) comes from the abstract. We also noticed that some individuals have very high numbers of variants (up to 36 in the exome). This may be a consideration for aggregate statistical analysis.

3.0.8 deLigt2012, PMID:23033978

The data from this paper comes from Table S3 and the number of trios (n = 100) comes from the abstract.
3.0.9  DeRubeis2014, PMID:25363760
The data from this paper comes from Supplementary Table 3. The proband count for trios (n = 2,270) comes from the main text and the removal of the 825 SSC trios for denominator counts is based on Supplementary Table 1. This results in 1,445 individuals.

3.0.10  Dimassi2015, PMID:26138355
The data from this paper comes from Table 1 and the count of n = 10 is from the abstract.

3.0.11  epi4k2013, PMID:23934111
The data from this paper comes from Supplementary Table 2 with sample names from Supplementary Table 1. The count of n = 264 comes from the abstract.

3.0.12  Fromer2014, PMID:24463507
The data from this paper comes from Supplementary Table 1. The number of n = 623 trios comes from that supplementary table legend.

3.0.13  GoNL, PMID:24974849
The data from this paper was downloaded from the website that was in the URLs section of the paper: http://www.nlgenome.nl/ The file at this website was named GoNL_DNMs.txt The number of n = 250 trios comes from the abstract.

3.0.14  Gulsuner2013, PMID:23911319
The data from this paper comes from Table S3. The number n = 105 comes from the results section. The number n = 84 comes from the main text.

3.0.15  Halvardson2016, PMID:27334371
The data from this paper comes from Supplementary Table 2. The number of n = 39 trios comes from the abstract.

3.0.16  Hashimoto2015, PMID:26582266
The data from this paper comes from Table 1. The number of n = 30 trios comes from the abstract.

3.0.17  Helbig2016, PMID:26795593
The data from this paper comes from Tables S3 and S7. The number n = 254 comes from the “Exome sequencing strategy” section of the paper.

3.0.18  Homsy2015, PMID:26785492
The data from this paper comes from Supplementary Database S2. The number n = 1,213 trios comes from the main text.

3.0.19  Iossifov, PMID:25363768
The data from this paper comes from Supplementary Table 2. The numbers 2,508 and 1,911 come from the main text.
3.0.20 Kataoka2016, PMID:27217147
The data from this paper comes from Table S2. The number n = 79 comes from the abstract.

3.0.21 Kranz2015, PMID:26091878
The data from this paper comes from Table 1. The number n = 14 comes from the abstract.

3.0.22 Krumm, PMID:25961944
The data from this paper comes from Supplementary Table 1. The number n = 2,377 comes from the abstract and the number 1,786 comes from the methods section.

3.0.23 Kun-Rodrigues, PMID:26362251
The data from this paper comes from Table 1. The number n = 21 comes from the abstract.

3.0.24 Lee2014, PMID:24501278
The data from this paper comes from Table 1. We only retained one of the monozygotic twin pairs for the database.

3.0.25 Lelieveld2016, PMID:27479843
The data from this paper comes from corrected Supplementary Table 2. The number n = 820 comes from the main text.

3.0.26 Lemay2015, PMID:25805808
The data from this paper comes from Table 1. The number n = 43 comes from the abstract.

3.0.27 Lifton, PMID:23665959
The data from this paper comes from Table S4. The number n = 362 comes from the abstract.

3.0.28 McCarthy2014, PMID:24776741
The data from this paper comes from Supplementary Table 3. The number n = 57 comes from the abstract.

3.0.29 Michaelson2012, PMID:23260136
The data from this paper comes from Table S1 and was lifted over to hg19. The number of 10 monozygotic twin pairs comes from Figure S1. Since variants from the Michaelson paper come from a set of 10 monozygotic twin pairs, only one set of each pair should be considered if doing large aggregate de novo statistics.

3.0.30 Moreno-Ramos2015, PMID:26352270
The data from this study comes from Table 1.

3.0.31 Ramu2013, PMID:23975140
The data from this study comes from Supplementary Table 9. The number n = 1 comes from the main text.
3.0.32 Rauch2012, PMID:23020937
The data from this study comes from Supplementary Tables S2, S3, S6, S7, and S8. The number n = 51 and n = 20 comes from the supplement under the “Study patients” section.

3.0.33 Rovelet-Lecrux 2015, PMID:26194182
The data from this paper comes from Table 2. The number n = 12 is from the abstract.

3.0.34 Sifrim2016, PMID:27479907
The data from this paper comes from Supplementary Tables 21, 22, and 23. The Sifrim et al. 2016 study on congenital heart disorders had overlap with other studies in denovo-db. In particular, it overlapped with the DDD 2017 and Lifton studies. We removed all samples from the Sifrim paper with a DDD in the name or a CG in the name that corresponded to these two studies. Supplementary Table 2 in Sifrim paper has the information for trios and based on that table there are 859 total trios in the Sifrim 2016 set (minus DDD and Lifton).

3.0.35 Slavotinek2015, PMID:25457163
The data from this paper were derived directly from the main text.

3.0.36 Smedemark-Margulies2016, PMID:27626066
The data from this paper comes from the main text under the “Genomic Analyses” section.

3.0.37 Smith2014, PMID:25105228
The data from this paper comes from the main text.

3.0.38 Steinberg2015, PMID:25773295
The data from this paper comes from Table 3.

3.0.39 Tavassoli2014, PMID:24650168
The data from this paper comes from the main text.

3.0.40 Turner2016, PMID:26749308
The data from this paper comes from Table S9. The number n = 53 comes from the abstract and the number n = 40 comes from the main text.

3.0.41 VanBon, PMID:22608503
The data from this paper comes from Table S3.

3.0.42 Veeramah2012, PMID:22365152
The data from this paper comes from the main text.

3.0.43 Veeramah2013, PMID:23647072
The data from this paper comes from Table 2.
3.0.44 Yu2015, PMID:26034137
The data from this paper comes from Table 1.

3.0.45 Yuen2016, PMID:27525107
The data from this paper comes from Supplementary Tables S4 and S5. The number $n = 200$ comes from the abstract.

3.0.46 Yuen2017, PMID:28263302
The data from this paper comes from Supplementary Tables S3, S4, and S5. The total number $n = 1,627$ comes from the main text. Note, there are two unaffected individuals with de novo variants based on Table S1.

4 References


