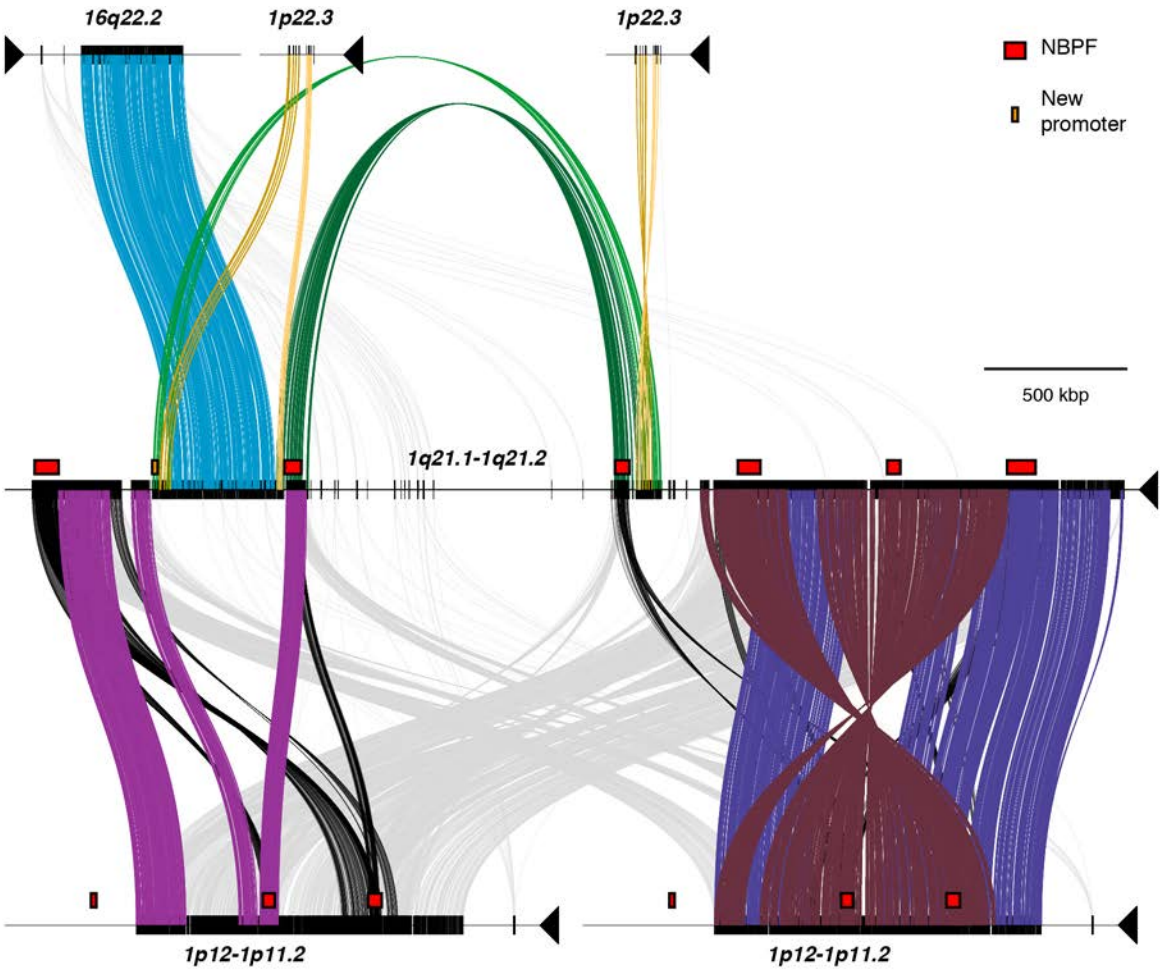
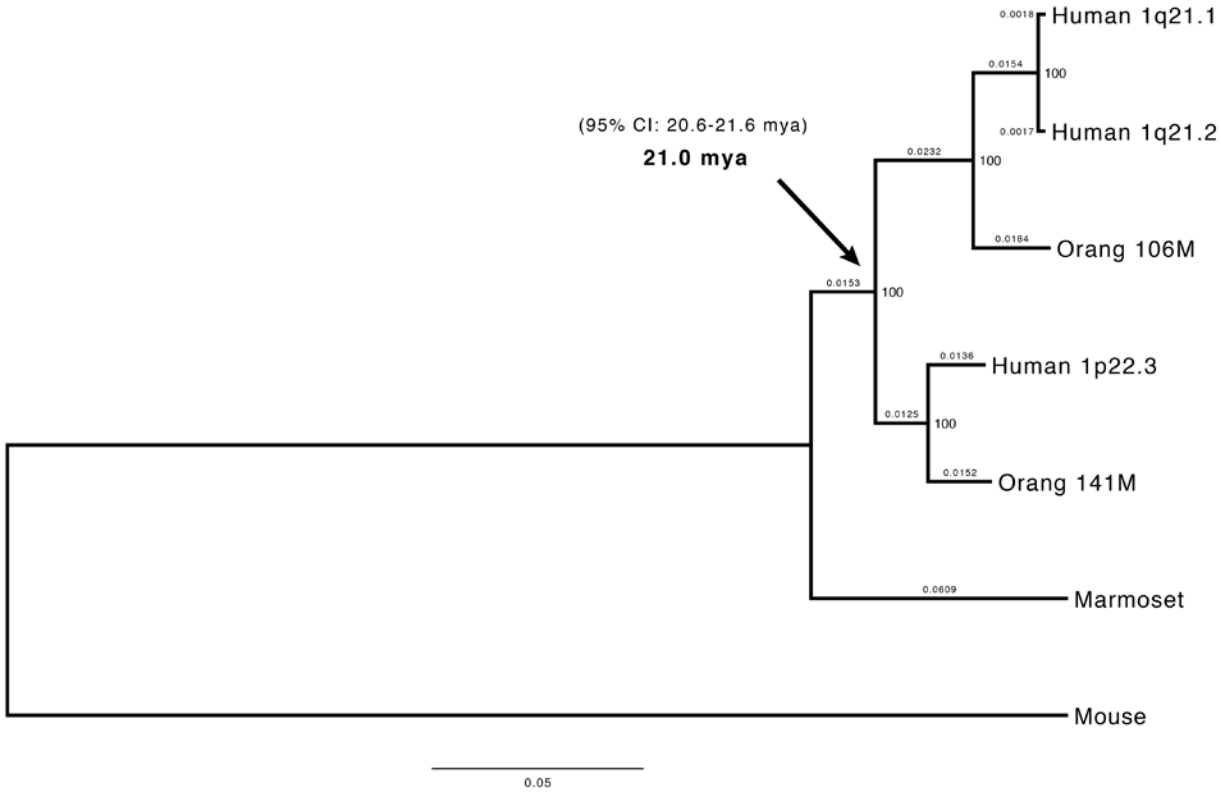
**ADDITIONAL FILE 1**




**Supplementary Figure 1. *HYDIN* and associated chromosome 1 duplications.** The segmental duplications that led to the formation of *HYDIN2* are seen in the context of larger genomic rearrangements on chromosome 1. Duplications are visualized using Miropeats (s = 800; Parsons, 1995). The central sequence represents chromosome 1q21.1-1q21.2 (chr1:146061572-150028860) with the *HYDIN* duplication in light blue. Also highlighted is the new *HYDIN2* promoter. The homology with chromosome 16q22.2 (chr16:70611384-71368670) and chromosome 1p22.3 (chr1:87315068-87609100) shown above the 1q21 region is the same as in Figure 1e. Shown below the 1q21 region is additional homology to chromosome 1p12-1p11.2 (chr1:119530046-121401465). This region is shown twice for clarity. Locations of the core duplicon gene *NBPF*, by GENCODE annotation, are highlighted as red boxes and are found at or near the breakpoints of most observed rearrangements, including the palindromic duplication shown at the bottom right (also described in O'Bleness, 2014).




**Supplementary Figure 2A. Timing of first duplication of the 5' and 3' segments flanking *HYDIN2*.** We estimate that the segmental duplication (65 kbp) that was bisected by the *HYDIN2* duplication underwent an earlier duplication from chromosome 1p22 to chromosome 1q21 approximately 21.0 mya (95% CI: 20.6-21.6 mya). This is consistent with the observation that this segment is at haploid single-copy in marmoset, macaque, and baboon, and at haploid copy number 2 in gibbon and orangutan.

*Sequences*. Sequences homologous to the three human loci were identified in primates and in mouse using BLAT and the UCSC Genome Browser. A single homologous locus was identified in mouse, which was used to root the tree. One homologous locus was identified in marmoset, representing New World monkeys, and one homologous locus was identified in both the rhesus macaque and baboon, representing Old World monkeys. This suggests the duplication occurred after the divergence of apes from Old World monkeys. Within apes, two homologous loci were identified in gibbon and orangutan each, suggesting that the duplication occurred within the common ancestor of hominoids (lesser and greater apes). Multiple homologous loci of varying lengths were identified in chimpanzee, suggesting this region has undergone further rearrangement and amplification in that lineage.

*Alignment and tree building*. A 30,872 bp multiple sequence alignment (MSA) was generated using MAFFT with sequences deriving from the single locus in mouse and marmoset, the two loci in orangutan, and the three loci in human. The phylogenetic tree was inferred using the maximum-likelihood method with distances estimated using the Tamura-Nei model and tested with 50 bootstrap replicates. Branches are labeled with the number of substitutions per site and nodes are labeled with bootstrap support.

*Timing the duplication*. The two loci on human 1q21 pass the relative rate test ($p = 0.65$, outgroup orangutan), as do either with their ortholog in orangutan ($p = 0.89$ for 1q21.1 and $p = 0.61$ for 1q21.2, outgroup marmoset). The locus on human 1p22.3 and its closest orangutan

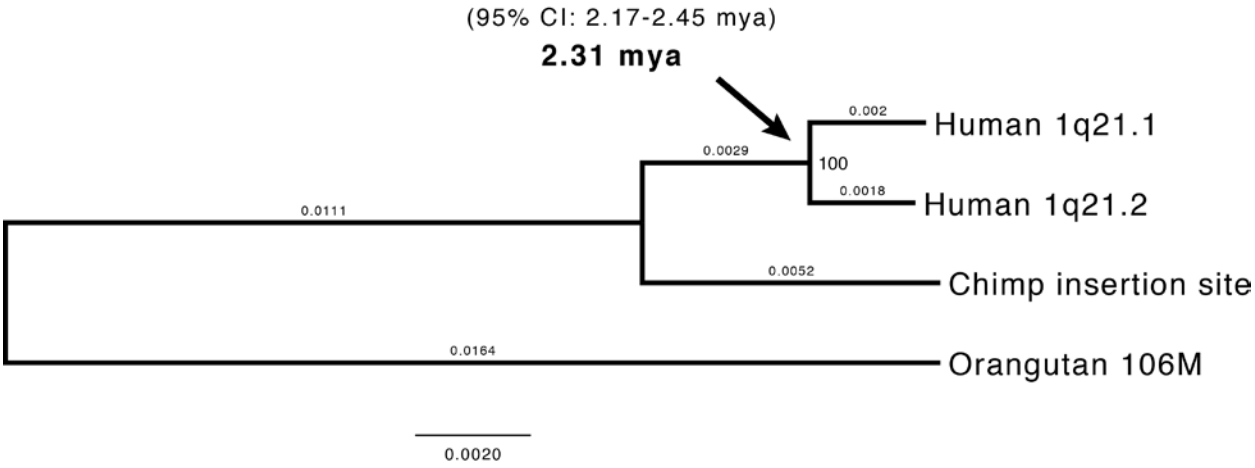
ortholog also pass the relative rate test ($p = 0.25$). However, the 1q and 1p clades appear to be evolving at different rates ($p < 0.00001$ for both, outgroup marmoset), and neither clade passes the relative rate test with marmoset ($p = 0.004$ for 1q and $p < 0.00001$ for 1p, outgroup mouse). In marmoset, the derived sequence sits on the long arm of chromosome 7, consistent with known translocations that differentiate the species (Sherlock, 1996). All in all, we cannot assume the same local rate of neutral substitution between any pair in these three clades (marmoset, 1p clade, 1q clade).
We use the 1p22.3 clade to estimate the timing of the duplication since that region is syntenic to the original sequence in marmoset. These estimates are based on the proportion of genetic distance since divergence from marmoset that occurred since the duplication of our segment of interest.

*clade 1p*
$D_{\text{human 1p22.3 branch}} = 0.013632$
$D_{\text{orangutan 141M branch}} = 0.015181$
$\text{Average}(D_{\text{orangutan 141M branch}}, D_{\text{human 1p22.3 branch}}) = 0.014407$
$D_{\text{1p until human-orang split}} = 0.012466$
$D_{\text{from divergence from marmoset to duplication}} = 0.015257$
(Distance after duplication)/(Total distance since divergence from marmoset) =
(0.014407 + 0.012466)/(0.014407 + 0.012466 + 0.015257) = 0.63785

Time of human–marmoset divergence (Glazko and Nei 2002) = 33 mya (range 32-36)

Estimate based on *clade 1p*
0.63785 * 33 mya = **21.0 mya** (95% CI: 20.6-21.6 mya)

We calculated 95% confidence intervals around our duplication timing estimate above using branch length error estimates and the following approach. First, for each branch in the tree, we set the branch length to a randomly chosen value between the actual branch length minus the branch length error (or zero if that value is negative) and the actual branch length plus the branch length error, inclusive. Second, we recomputed the timing estimate above using the same calculations as for the original tree except using the modified branch length values. For these calculations, we assumed a human–marmoset divergence of 33 mya. Third, we repeated the above two steps until we obtained one million modified trees and corresponding timing estimates. Finally, we sorted the estimates and reported the $25{,}000^{th}$ and $975{,}000^{th}$ sorted timing estimate values as the 95% confidence interval around the corresponding timing point estimate: 20.6-21.6 mya. Note that this error is less than the error that results from uncertainty in the timing estimate for human–marmoset divergence (Glazko and Nei, 2003).
This range of estimates, which places the duplication just after the divergence of apes and Old World monkeys (23 mya, range 21-25), is consistent with where we place the duplication in evolutionary history based on its presence and absence in reference genomes. It is important to note, however, that nonhuman primate reference genomes may be incomplete, especially in these duplicated regions.

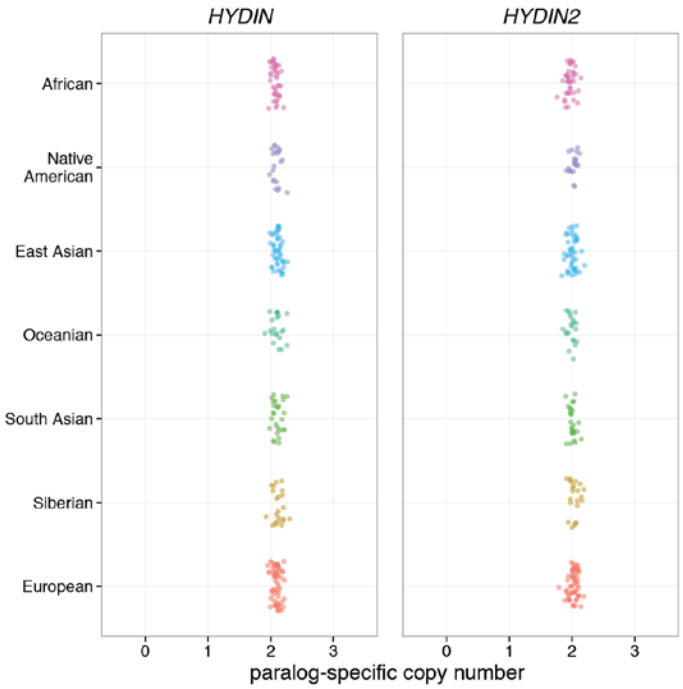



**Supplementary Figure 2B. Timing of second duplication of the 5' and 3' segments flanking *HYDIN2*.** To time the more recent duplication of the sequence immediately flanking *HYDIN2*, which occurred between 1q21.1 and 1q21.2, a new MSA was generated with sequence from the two human loci, the homologous chimpanzee locus, and the homologous orangutan locus. The sequences for the two human loci and the orangutan locus are the same as used in Figure S2a. Because this region appears to have undergone further duplication in chimpanzee (data not shown), we identified and sequenced a BAC containing homologous chimpanzee sequence (CH251-231E10). A 209,299 bp MSA was generated using MAFFT and manually edited for obvious alignment errors. The phylogenetic tree was inferred using the maximum-likelihood method with distances estimated using the Kimura 2-paramter model and tested with 50 bootstrap replicates. Branches are labeled with the number of substitutions per site and nodes are labeled with bootstrap support.

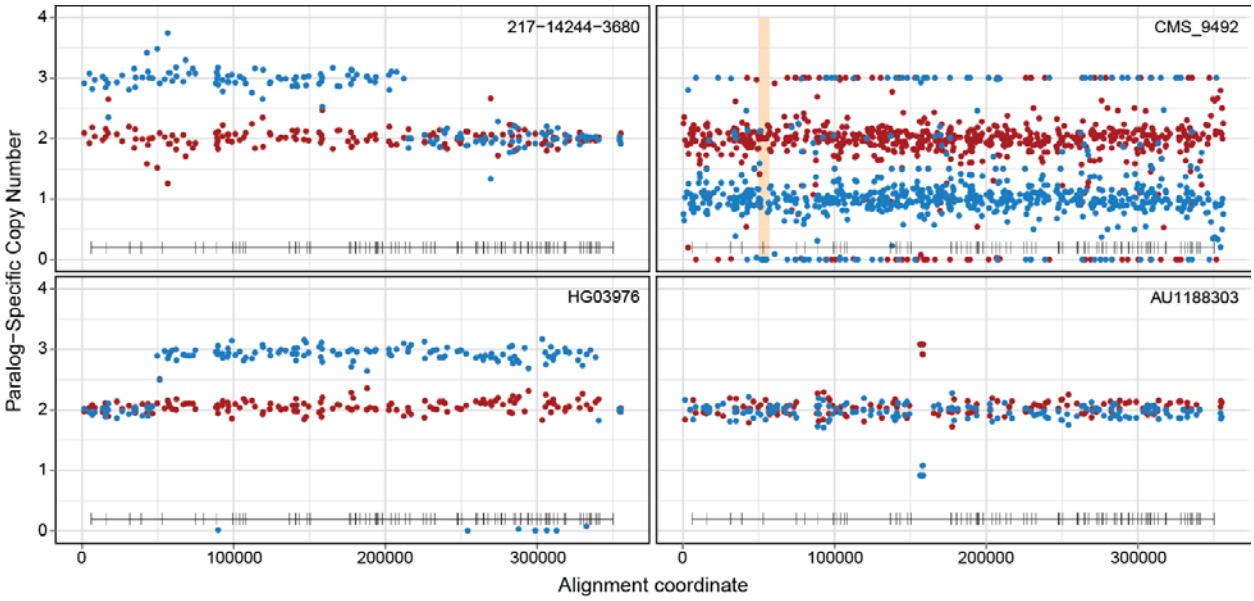
*Timing the human-specific duplication*. We similarly estimate the timing of the human-specific duplication of the segments that flank *HYDIN2* using orangutan as the outgroup, and a divergence time of 6 mya (Glazko and Nei, 2003), based on the proportion of genetic distance since divergence from chimpanzee that occurred since the duplication of our segment of interest. The human branches pass the relative rate test ($p = 0.18$).

$D_{\text{average human 1q21.1/1q21.2 branch}} = (0.002032 + 0.001846)/2 = 0.001939$
$D_{\text{1q human until 1q21.1/1q21.2 split}} = 0.002932$
$D_{\text{1q human total}} = (0.002032 + 0.001846)/2 + 0.002932 = 0.004871$
(Distance after duplication)/(Total evolutionary distance since divergence from chimpanzee) = 0.001939/(0.004871 + 0.005213) = 0.1923
0.1923 * 2 * 6 mya = **2.31 mya** (95% CI: 2.17-2.45 mya)*

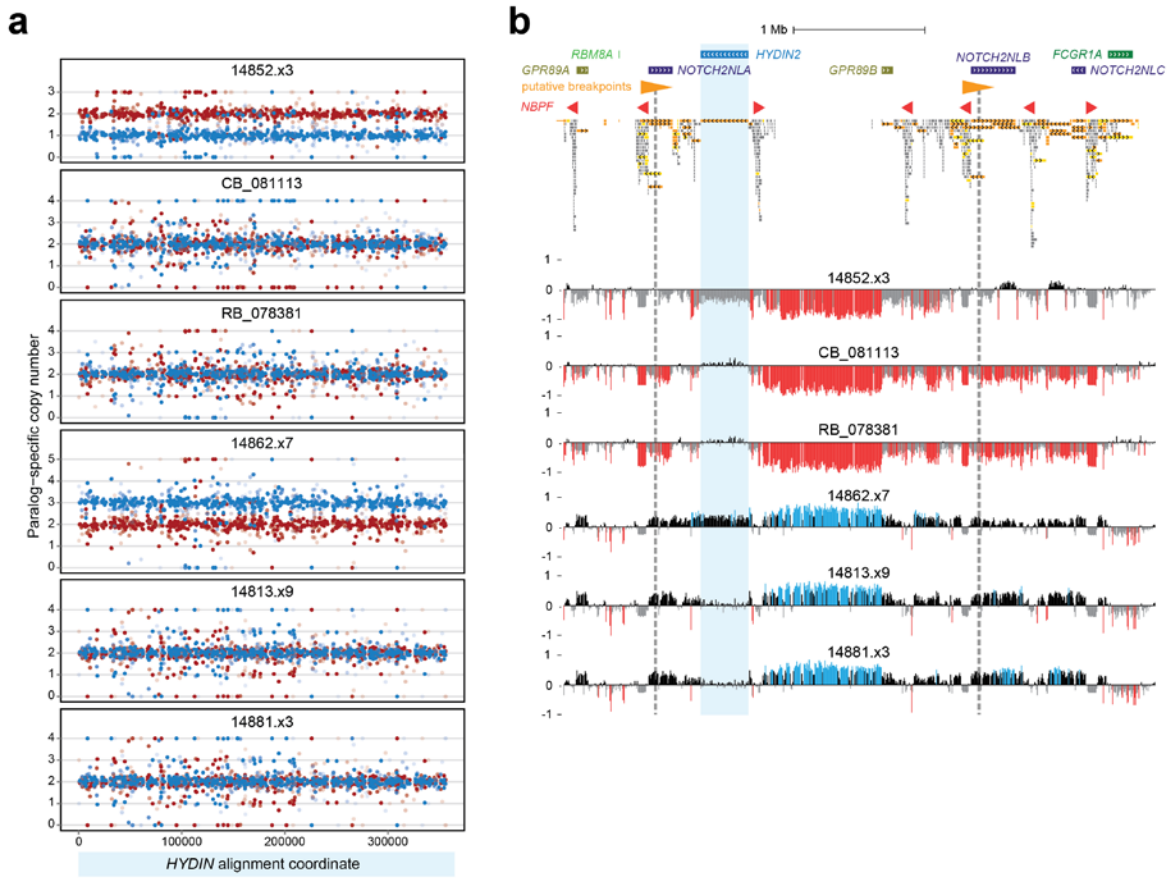
This estimate is more recent than the estimate of the *HYDIN2* duplication, however it must necessarily have preceded the insertion of the duplicated *HYDIN2* sequence. We think it most likely that interlocus gene conversion is responsible for reducing the genetic divergence between these segments and distorting this timing estimate.



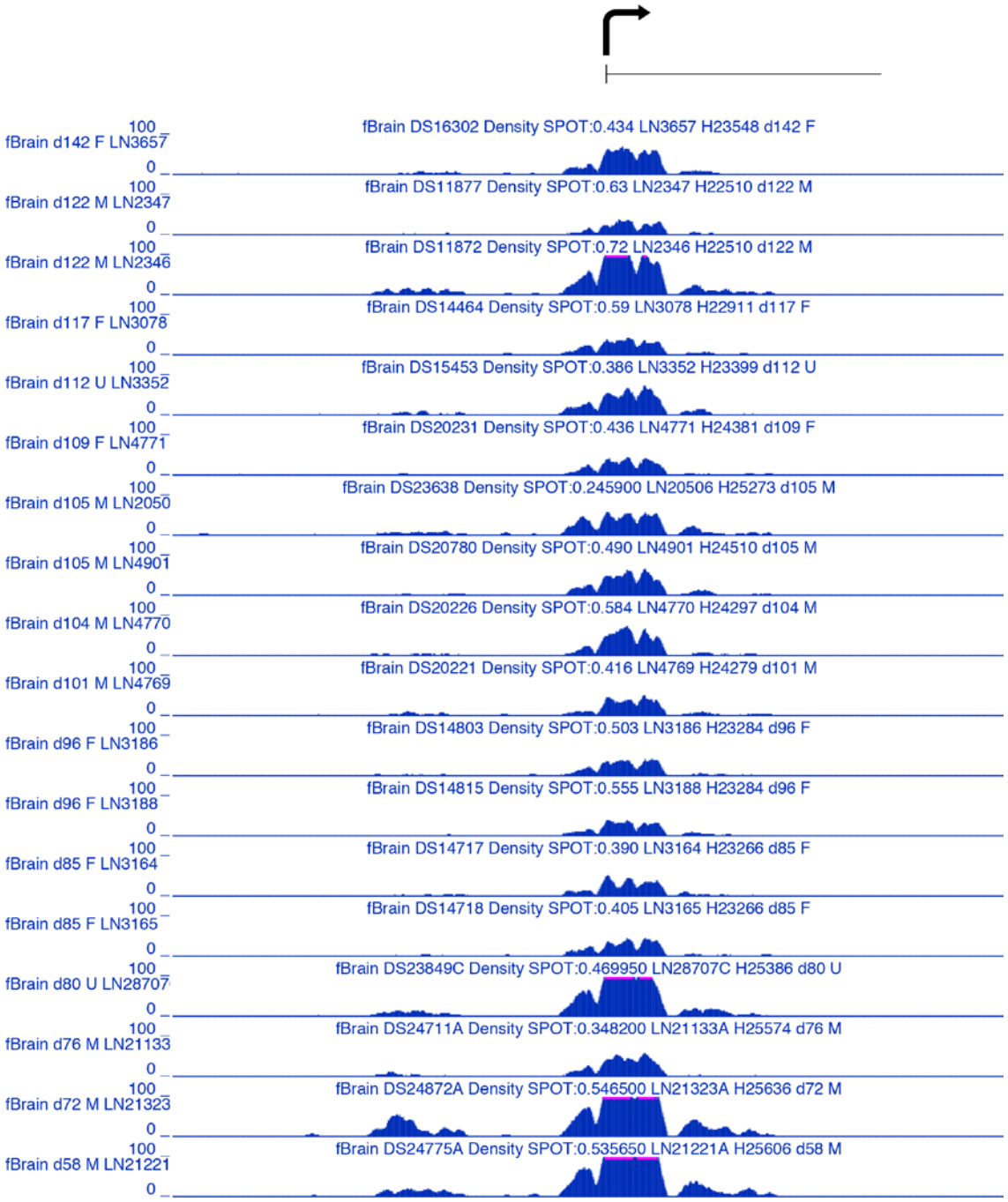

**Supplementary Figure 3. Paralog-specific copy number estimates for 236 individuals from the Human Genome Diversity Project (HGDP).** Whole-genome sequencing data from the HGDP were mapped to the *HYDIN* segmental duplication, and SUNK-based read depth was used to assess paralog-specific diploid copy number. Both paralogs are found at copy number 2 in all individuals shown, including 41 Africans, 21 Native Americans, 45 East Asians, 21 Oceanians, 27 South Asians, 22 Siberians, and 59 Europeans.




**Supplementary Figure 4. *HYDIN* internal structural variation and interlocus gene conversion.** Examples of structural variation within *HYDIN* paralogs and a putative interlocus gene conversion event. Each point shows a paralog-specific copy number estimate (red, *HYDIN*; blue, *HYDIN2*) based on sequencing data corresponding to a single MIP targeting sequence that distinguishes *HYDIN* paralogs. Sequencing reads were analyzed to compute paralog-specific read count relative frequencies for each MIP, which were multiplied by the aggregate estimated copy number at each target site to infer paralog-specific copy number. Shown are an ~212 kbp duplication affecting *HYDIN2* (upper left), an ~289 kbp duplication affecting *HYDIN2* (lower left), a putative ~3 kbp deletion affecting *HYDIN2* in a 1q21 microdeletion patient (orange highlight, upper right), and an ~2 kbp interlocus gene conversion event (lower right). Note that the putative *HYDIN2* deletion shown might instead reflect interlocus gene conversion—all reads for 11 consecutive MIPs over the highlighted interval mapped to *HYDIN*, consistent with zero copies of *HYDIN2* and either two or three copies of *HYDIN*. Also note that the interlocus gene conversion event identified in AU1188303 is the same as that discovered in 11094.s1 (**Figure 2C**, bottom right), indicating this event likely segregates at very low frequency. 153 MIPs were used for genotyping all individuals shown except CMS_9492, a 1q21 microdeletion patient genotyped with 717 MIPs. All events shown were detected by an automated caller. Duplicated exons based on the canonical *HYDIN* gene model are indicated at the bottom of each plot.




**Supplementary Figure 5. 1q21 rearrangement breakpoint variability. a)** 717 MIPs targeting regions that distinguish *HYDIN* paralogs were employed to genotype *HYDIN* paralog-specific copy number in 48 1q21 microdeletion and 25 1q21 microduplication patients. Points show *HYDIN* paralog-specific copy number estimates (red, *HYDIN*; blue, *HYDIN2*) for three microdeletion patients (14852.x3, CB_081113, and RB_078381) and three microduplication patients (14862.x7, 14813.x9, and 14881.x3). These estimates were calculated as the product of the paralog-specific read count relative frequency for a particular MIP and the aggregate estimated copy number at the corresponding target site. The MIP results indicate that 1q21 rearrangements do not always include *HYDIN2*. **b)** The segmental duplication organization of a 4.5 Mbp region at chromosome 1q21 (GRCh38 chr1:145,500,001-150,000,000) is shown along with array CGH profiles for the individuals in panel a. Thin colored boxes indicate sequences duplicated between this region and another genomic locus, with colors corresponding to sequence identity (orange = 99% or above, yellow = 98%–99%, gray = 90%–98%) and markings showing orientation (right-pointing, directly oriented; left-pointing, inversely oriented) between duplication pairs. Thick colored boxes highlight locations of several duplicated genes. Orange triangles indicate high-identity, directly oriented *NOTCH2NL-NBPF* duplications, with putative breakpoints of the canonical 1q21 rearrangement shown as vertical gray dashed lines. Shown below are the locations of *NBPF* core duplicons. Array CGH confirms 1q21 rearrangements in these individuals, and array data over *HYDIN2* (blue highlight) indicate a loss or gain in some (14852.x3 and 14862.x7) but not others (all other individuals shown), validating the MIP data shown in panel a.




**Supplementary Figure 6. The *HYDIN2* promoter corresponds to a peak of chromatin accessibility in fetal brain.** Reads indicating sites of chromatin accessibility as determined by sensitivity to DNase I digestion from various fetal brain time points (day 58 – day 142) were mapped using mrsFAST-Ultra (Hach, 2014) to allow for measurement over duplicate sequence. Shown is a ~14 kbp region (chr1:146479496-146493419) surrounding the acquired promoter and first exon of *HYDIN2*, with a peak is visible at all time points. See Table S8 for sample information.

**Supplementary Table 1. MIP-based copy-number genotyping of *HYDIN2*.**

Number of individuals

| Cohort | Individuals genotyped | Cases or controls | *HYDIN2* deletions (CN = 1) | *HYDIN2* normal (CN = 2) | *HYDIN2* duplications (CN = 3) | *HYDIN2* complex event |
|---|---|---|---|---|---|---|
| SVIP 16p | 336 | both | 0 | 336 | 0 | 0 |
| *16p triplications* | *1* | | *0* | *1* | *0* | *0* |
| *16p duplications* | *70* | | *0* | *70* | *0* | *0* |
| *16p normals* | *187* | | *0* | *187* | *0* | *0* |
| *16p deletions* | *78* | | *0* | *78* | *0* | *0* |
| SVIP 1q21 | 120 | both | 45 | 52 | 22 | 1 |
| *1q21 duplications* | *25* | | *0* | *3* | *22* | *0* |
| *1q21 normals* | *46* | | *0* | *46* | *0* | *0* |
| *1q21 deletions* | *49* | | *45* | *3* | *0* | *1* |
| HapMap/1KG | 1082 | controls | 1 | 1076 | 4 | 1 |
| AGRE Probands | 941 | cases | 0 | 941 | 0 | 0 |
| SSC Probands | 787 | cases | 0 | 786 | 0 | 1 |
| TASC Probands | 890 | cases | 0 | 888 | 1 | 1 |
| TASC Siblings | 123 | controls | 0 | 123 | 0 | 0 |
| AGRE Siblings | 638 | controls | 0 | 637 | 0 | 1 |
| SSC Siblings | 1133 | controls | 0 | 1130 | 1 | 2 |
| SSC Parents | 2 | controls | 0 | 2 | 0 | 0 |
| Cell Lines | 3 | controls | 0 | 3 | 0 | 0 |
| All Combined | 6055 | both | 46 | 5974 | 28 | 7 |
| *Cases (excluding SVIP)* | *2618* | *cases* | *0* | *2615* | *1* | *2* |
| *Controls (excluding SVIP)* | *2981* | *controls* | *1* | *2971* | *5* | *4* |
| *SVIP cohorts* | *456* | *both* | *45* | *388* | *22* | *1* |

Number of unrelated chromosomes

| Cohort | Chromosomes genotyped | Cases or controls | *HYDIN2* deletions (CN = 1) | *HYDIN2* normal (CN = 2) | *HYDIN2* duplications (CN = 3) | *HYDIN2* complex event |
|---|---|---|---|---|---|---|
| HapMap/1KG_YRI | 117/117 | controls | 1/1 | 115/115 | 1/1 | 0/0 |
| HapMap/1KG_GBR | 185/186 | controls | 0/0 | 185/186 | 0/0 | 0/0 |
| HapMap/1KG_FIN | 190/190 | controls | 0/0 | 190/190 | 0/0 | 0/0 |
| HapMap/1KG_CHS | 4/4 | controls | 0/0 | 4/4 | 0/0 | 0/0 |
| HapMap/1KG_CHB | 96/96 | controls | 0/0 | 96/96 | 0/0 | 0/0 |
| HapMap/1KG_PUR | 4/4 | controls | 0/0 | 4/4 | 0/0 | 0/0 |
| HapMap/1KG_IBS | 190/190 | controls | 0/0 | 190/190 | 0/0 | 0/0 |
| HapMap/1KG_KHV | 10/10 | controls | 0/0 | 10/10 | 0/0 | 0/0 |
| HapMap/1KG_CDX | 2/2 | controls | 0/0 | 2/2 | 0/0 | 0/0 |
| HapMap/1KG_GWD | 10/10 | controls | 0/0 | 10/10 | 0/0 | 0/0 |
| HapMap/1KG_PJL | 2/2 | controls | 0/0 | 2/2 | 0/0 | 0/0 |
| HapMap/1KG_ESN | 10/10 | controls | 0/0 | 10/10 | 0/0 | 0/0 |
| HapMap/1KG_MSL | 6/6 | controls | 0/0 | 6/6 | 0/0 | 0/0 |
| HapMap/1KG_ITU | 8/8 | controls | 0/0 | 7/7 | 0/0 | 1/1 |
| HapMap/1KG_CEU | 237/243 | controls | 0/0 | 237/243 | 0/0 | 0/0 |
| HapMap/1KG_JPT | 102/102 | controls | 0/0 | 102/102 | 0/0 | 0/0 |
| HapMap/1KG_LWK | 140/151 | controls | 0/0 | 140/151 | 0/0 | 0/0 |
| HapMap/1KG_ASW | 115/131 | controls | 0/0 | 113/129 | 2/2 | 0/0 |
| HapMap/1KG_MXL | 112/115 | controls | 0/0 | 112/115 | 0/0 | 0/0 |
| HapMap/1KG_TSI | 174/174 | controls | 0/0 | 174/174 | 0/0 | 0/0 |
| HapMap/1KG_GIH | 2/2 | controls | 0/0 | 2/2 | 0/0 | 0/0 |
| HapMap/1KG_MKK | 150/150 | controls | 0/0 | 150/150 | 0/0 | 0/0 |
| HapMap/1Kg_all_combined | 1866/1903 | controls | 1/1 | 1861/1898 | 3/3 | 1/1 |

*First number (before slash) is minimum number of unrelated chromosomes, second number (after slash) is maximum number of unrelated chromosomes.

**Supplementary Table 2. *HYDIN* duplication, deletion, and interlocus gene conversion events.**

| Variant* | Carrier(s) | Carrier status(es) | Called by | Alignment location | Genic location** | Minimum size | Number of supporting MIPs | Transmission observed?† | Previously known? |
|---|---|---|---|---|---|---|---|---|---|
| *HYDIN* internal duplication | 03C16388 | case | automated caller | 225785-240797 | intron 42 to intron 45 | 15 kbp | 9 | no | no |
| *HYDIN2* internal deletion | CMS_9492 | case (chromosome 1q21 deletion) | automated caller | 51702-55192 | intron 9 to intron 10 | 3 kbp | 11 | no | no |
| *HYDIN2* internal deletion | SSC 11176.s1 | control | automated caller | 138752-141883 | intron 19 to intron 20 | 3 kbp | 5 | no | no |
| *HYDIN2* internal duplication | 217-14244-3680 | case | automated caller | 1-212135 | intron 5 to intron 39 | 212 kbp | 92 | no | no |
| *HYDIN2* internal duplication | HG03976 | control | automated caller | 49583-338591 | intron 9 to intron 80 | 289 kbp | 129 | no | no |
| *HYDIN2* internal duplication | SSC 11422.p1 | case | automated caller | 54502-212135 | intron 10 to intron 39 | 158 kbp | 70 | no | no |
| *HYDIN -> HYDIN2* conversion | SSC 11094.s1, AU1188303 | control, control | automated caller | 156380-158432 | intron 23 | 2 kbp | 5 | no | no |
| *HYDIN* deletion | SSC Si1295 | case | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 272 | no | yes |
| *HYDIN* duplication | 05C47106 | case | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 153 | no | yes |
| *HYDIN* duplication | Ag449 | case | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 272 | no | yes |
| *HYDIN2* deletion | NA19190 | control | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 272 | no | yes |
| *HYDIN2* duplication | NA19201 | control | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 272 | yes | yes |
| *HYDIN2* duplication | NA19703 | control | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 272 | yes | yes |
| *HYDIN2* duplication | NA19705 | control | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 153 | yes | yes |
| *HYDIN2* duplication | NA20127 | control | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 272 | yes | yes |
| *HYDIN2* duplication | 215-13135-1523 | case | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 153 | yes | yes |
| *HYDIN2* duplication | SSC 12325.s2 | control | automated caller | 1-357274 | intron 5 to intron 84 | 357 kbp | 153 | yes | yes |

*For interlocus gene conversion variants, the conversion donor is listed before the arrow and the conversion acceptor after the arrow.

**Exons and introns are numbered according to the *HYDIN* gene model, with exons 6-84 shared between *HYDIN* and *HYDIN2*.

†Did we observe at least one instance of transmission of this variant within a trio? Note that for most individuals, DNA from parents and/or children was not available, so not observing transmission does not necessarily indicate the variant is *de novo*.

**Supplementary Table 3: Locations of other copies of the *HYDIN2* promoter-associated duplication, their relationship to *NBPF*, and evidence for transcription**

| BLAT result | chr | start | end | % identity | associated *NBPF* member | distance to EV15 promoter (kbp) | distance to CM promoter (kbp) | distance to *NBPF* 3' UTR | orientation (promoter/ NBPF) | relationship of promoter to *NBPF* | spliced ESTs* | tissue source(s) of spliced EST(s)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | chr1 | 146427435 | 147030676 | 100.0% | *NBPF12* | 451 | 472 | | +/+ | upstream | yes | hippocampus (3), brain (1), fetal brain (1) |
| 2 | chr1 | 148088687 | 148263630 | 99.5% | *NBPF11* | 98 | 118 | | -/- | upstream | yes | fetal brain (1) |
| 3 | chr1 | 145239245 | 145540744 | 91.5% | *NBPF20* | 49 | 69 | | +/- | upstream | yes | hippocampus (1) |
| 4 | chr1 | 148002800 | 148304299 | 91.4% | *NBPF11* | | | 52 | +/- | downstream | yes | hippocampus (1), NE lung carcinoid (2), carcinoid (2) |
| 5 | chr1 | 145973900 | 146728039 | 90.8% | *NBPF10* | | 222 | | -/- | upstream | yes | cerebellum (2) |
| 6 | chr1 | 148471458 | 148681704 | 90.2% | *NBPF14* | | | 2.1 | -/- | downstream | yes | pooled tissues (1) |
| 7 | chr1 | 144393920 | 144469916 | 90.3% | *NBPF15* | | | 2.1 | -/- | downstream | no | |
| 8 | chr1 | 145239245 | 145540744 | 90.2% | *NBPF20* | | | 2.1 | -/- | downstream | no | |
| 9 | chr1 | 149354590 | 149691092 | 90.6% | *NBPF19* | | | 2.1 | -/- | downstream | yes | germinal center B-cell (1), embryonic 1st PA (1) |
| 10 | chr1 | 120795255 | 120894554 | 90.0% | *NBPF26* | | | 2.8 | +/+ | downstream | yes | brain (1), spleen (1) |

*Are spliced human ESTs present with 5' ends contained within the promoter?

**When >5 spliced ESTs were present, only tissue sources for 5 were recorded, chosen by order of display on the UCSC genome browser. NE: neuroendocrine; PA: pharyngeal arch.

**Supplementary Table 4. Pairwise dN/dS values for *HYDIN* in primates.**

| Species/Paralog | Human duplicate | Human ancestral | Chimpanzee | Gorilla | Gibbon | Macaque | Marmoset |
|---|---|---|---|---|---|---|---|
| Human duplicate | - | | | | | | |
| Human ancestral | 0.47 | - | | | | | |
| Chimpanzee | 0.39 | 0.29 | - | | | | |
| Gorilla | 0.28 | 0.23 | 0.24 | - | | | |
| Gibbon | 0.29 | 0.28 | 0.28 | 0.26 | - | | |
| Macaque | 0.25 | 0.24 | 0.25 | 0.23 | 0.26 | - | |
| Marmoset | 0.31 | 0.30 | 0.30 | 0.30 | 0.31 | 0.31 | - |

**Supplementary Table 5: Likely gene disruptive events detected in *HYDIN/HYDIN2* by MIP-based sequencing of exons.**

| Paralog* | Variant | Exon | Intron | Protein position | Amino acid | Samples | Frequency | Number genotyped** |
|---|---|---|---|---|---|---|---|---|
| *HYDIN* | splice_donor | - | 29/85 | - | - | 1 | 0.04% | 2603 |
| *HYDIN* | stop_gained | 11/86 | 11/86 | 1330 | R/* | 1 | 0.04% | 2599 |
| *HYDIN2* | splice_acceptor | - | 14/85 | - | - | 2 | 0.08% | 2605 |
| *HYDIN2* | frameshift | 19/86 | - | 2531-2532 | A/X | 1 | 0.04% | 2607 |
| *HYDIN2* | splice_donor | - | 28/85 | - | - | 1 | 0.04% | 2600 |
| *HYDIN2* | frameshift | 41/86 | - | 2115-2116 | VI/VSX | 10 | 0.38% | 2598 |
| *HYDIN2* | splice_donor | - | 42/85 | - | - | 1 | 0.04% | 2599 |
| *HYDIN2* | frameshift | 46/86 | - | 2485 | A/X | 1 | 0.04% | 2600 |
| *HYDIN2* | frameshift | 48/86 | - | 2680 | G/X | 1 | 0.04% | 2598 |
| *HYDIN2* | splice_acceptor | - | 54/85 | - | - | 1 | 0.04% | 2595 |
| *HYDIN2* | splice_acceptor | - | 67/85 | - | - | 6 | 0.23% | 2601 |
| *HYDIN2* | stop_gained | 80/86 | - | 4563 | W/* | 1 | 0.04% | 2599 |
| unknown | splice_donor | - | 8/85 | - | - | 1 | 0.04% | 2601 |
| unknown | frameshift | 17/86 | - | 766-768 | LVL/X | 1 | 0.04% | 2597 |
| unknown | splice_donor | - | 31/85 | - | - | 1 | 0.04% | 2595 |
| unknown | frameshift | 35/86 | - | 1771 | P/X | 24 | 0.92% | 2595 |
| unknown | frameshift | 37/86 | - | 1885 | N/X | 1 | 0.04% | 2603 |
| unknown | splice_donor | - | 43/85 | - | - | 3 | 0.12% | 2602 |
| unknown | stop_gained | 45/86 | - | 2352 | R/* | 1 | 0.04% | 2597 |
| unknown | splice_donor | - | 50/85 | - | - | 1 | 0.04% | 2598 |
| unknown | frameshift | 51/86 | - | 2876 | T/NX | 1 | 0.04% | 2604 |
| unknown | stop_gained | 52/86 | - | 2928 | Q/* | 1 | 0.04% | 2595 |
| unknown | splice_acceptor | - | 52/85 | - | - | 5 | 0.19% | 2594 |
| unknown | splice_donor | - | 53/85 | - | - | 2 | 0.08% | 2601 |
| unknown | stop_gained | 62/86 | - | 3504 | E/* | 1 | 0.04% | 2604 |
| unknown | stop_gained | 70/86 | - | 3970 | R/* | 1 | 0.04% | 2604 |
| unknown | splice_donor | - | 75/86 | - | - | 1 | 0.04% | 2602 |
| unknown | splice_acceptor | - | 83/85 | - | - | 1 | 0.04% | 2598 |
| unknown | stop_gained | 84/86 | - | 4846 | Q/* | 1 | 0.04% | 2602 |

*Paralog determined by presence of SUN on variant-containing MIP reads, variants identified by MIP reads that did not intersect a SUN could not be assigned; Variants in *HYDIN2* are annotated with the exon numbering scheme from *HYDIN*

**Number of samples successfully genotyped for this variant (Freebayes)

**Supplementary Table 6. Phenotypes for patients having atypical chromosome 1q21 rearrangements.**

| Individual | Cohort | 1q21 CNV | *HYDIN2* copy number (MIP) | Phenotype | Notes |
|---|---|---|---|---|---|
| 14813.x10 | SVIP | duplication | 2 | The patient is a 6-year-old Caucasian male. Facial features include macrocephaly, mild bilateral down slanting palpebral fissures and small ears with thickened/overfolded helices and large lobes that are posteriorly rotated. Additional facial features include mild plagiocephaly, small mouth, and tented upper lip. Physical examination reveals left 5th and right 4th-5th clinodactyly of fingers as well as 4th and 5th bilateral clinodactyly of toes. Patient has an anterior hair whorl, a hypopigmented spot on his right abdomen and a Café-au-lait colored macule on his left inner knee. Examination also reveals lordosis and double sacral dimples with gluteal cleft at midline (shallow, but easily visualized). Patient currently has a significantly above average head circumference measurement (72 months: HC = 56.1, z = 2.68). Patient has a significantly above average height and weight, but has a normal BMI of 21.4 (72 months: height = 127.5 cm, z = 2.36 weight = 34.74 kg, z = 2.83). Patient was diagnosed with Autistic Disorder (confirmed with ADOS, ADI, and clinical judgment using DSM-IV criteria) as well as Mild Mental Retardation (based on diagnostic history, cognitive and adaptive assessment, parent report of symptoms and clinical judgment). He uses mostly single words with occasional phrases. Patient shows autism-related impairments in language and social communication, including odd intonation/pitch, echolalia, stereotyped language, limited spontaneous use of gestures and eye contact, limited range of facial expressions, limited response to social smile and name, difficulties with joint attention, initiation of social interaction and pretend play. Finger flapping and repetitive use of play objects are also noted. Patient's cognitive abilities fall in the Very Low range (Mullen Verbal IQ = 38, DAS-II Nonverbal IQ = 51) Adaptive abilities fall in the Low range (Vineland Adaptive Composite = 64). Patient used his first single words at 54 months of age. Earlier in his history, he did spontaneously sing/babble several simple songs, but mother reports loss of babbling/singing between ages 2 and 3. Abnormalities were first noted in his development at 24 months of age. Patient has low receptive and expressive communication (Vineland Communication Domain Standard Score = 63; Mullen Receptive Language Subscale Age Equivalent = 28 months, Mullen Expressive Language Subscale Age Equivalent = 27 months). Patient is right-hand dominant. Patient evidences moderate impairments in gross motor functioning and manual dexterity, per parent report (Vineland Motor Skills Domain Standard Score = 72). Patient's parent endorses concerns with attention problems and hyperactivity. Patient was born at 39 weeks gestation via cesarean section due to mother's anal fissures, but no other labor complications were reported. Following birth, patient experienced hyperbilirubinemia but did not receive treatment. Patient was diagnosed with intermittent constipation at 3 months of age, which has not been resolved. At 19 months of age, patient was diagnosed with heart murmurs and has a history of chronic pneumonia. In addition, patient was diagnosed with Tics at 2 years of age. Patient has strabismus, and was initially treated with glasses (later unnecessary). Parent reports that the patient is excessively clumsy and uncoordinated, but no other neurological diagnoses have been made. | |
| 14813.x9 | SVIP | duplication | 2 | The patient is a 37-year-old Caucasian male. Physical examination reveals pupillary hippus bilaterally, as well as an irregular bordered light brown macule on the left cheek. Patient has an irregular bordered light brown macule on his right upper back, as well as an oval café au-lait spot on his left buttock. Patient has wide-spaced 1st-2nd toes bilaterally, and very mild left concave thoracic scoliosis. Patient also has mild plagiocephaly. Additionally, he has a mild left deviation of the gluteal cleft at the top, but no dimple. Patient has a head circumference of 60.0cm. Patient has a height of 177.5cm and a weight of 122.92kg, with a BMI indicative of obesity. Patient has a diagnosis of Depressive Disorder Not Otherwise Specified (confirmed with rating scales, clinical judgment using DSM-IV criteria) for which he takes Sertraline. Patient does not meet diagnostic criteria for ASD or other related disorders, but is observed to have a flat affect and flat/monotone speech. He self-reports some social difficulties, few close friendships, and has moderately low scores on Vineland Socialization scale (Standard Score = 85). As a child he notes he was shy, stayed to himself and did not talk much. Patient's cognitive abilities fall in the Average range (WASI Verbal IQ = 111, Nonverbal IQ = 97, Full Scale IQ = 105). Besides the aforementioned scores on Socialization, his adaptive abilities fall in the Adequate range (Adaptive Composite = 90). Academic skills are Average to High Average (WIAT-III Reading Comprehension = 120; Sentence Composition = 92; Word Reading = 114; Numerical Operations = 106). Patient is right-hand dominant. Patient has Low Average fine motor coordination (Purdue Pegboard T scores, Dominant = 40.35, Non-dominant = 49.07, Both Hands = 37.68). Patient was born vaginally at full-term (exact gestational weeks unknown). No pregnancy or labor complications were reported. Patient wears glasses to correct vision to normal. Patient has macrocephaly, but no other neurological problems are noted. Patient has a structural intestinal malrotation. He has had several surgeries, including gallbladder removal, an appendectomy and an orchiopexy (for undescended testicles in childhood). Patient also has severe sleep apnea, and uses a CPAP machine. | father of 14813.x10 |
| 14881.x3 | SVIP | duplication | 2 | Patient is a 36-year old male. Physical examination reveals presence of a café au lait (location not specified). Patient also exhibits rapid and fast postural tremor. Patient currently meets criteria for macrocephaly (435 months: HC = 59.2 cm, z = 2.86). Patient has an above average height, an above average weight (435 months: height = 186.4 cm, z = 1.34, weight = 91 kg, z = 1.40) and a BMI indicative of being overweight (BMI = 26.2). Patient does not meet diagnostic criteria for ASD or other psychiatric disorders. Patient's cognitive abilities fall in the Average to Above Average ranges (WASI Verbal IQ = 107, Nonverbal IQ = 120, Full Scale IQ = 115). His adaptive abilities fall in the Average range (Adaptive Composite = 107). Patient is right-hand dominant and has Low Average fine motor coordination (Purdue Pegboard T scores, Dominant = 40, Non-dominant = 38, Both Hands = 38). Patient was born vaginally at 41 weeks gestation. While labor complications were endorsed, no specifics were obtained. Patient was diagnosed with recurrent otitis media (> 8 total occurrences) at 8 years of age, eczema at 10 years of age, and hypertension at 10 years of age. Additionally, patient was diagnosed with heart problems at 28 years of age (no specifics available) and intermittent pneumonia at 32 years of age, which has resolved. Patient was diagnosed with head injury/loss of consciousness at 10 years of age and macrocephaly at 15 years of age, but no other neurological problems are noted. No gastrointestinal conditions or sleep problems were reported | |
| CB_081113 | Manchester | deletion | 2 | Patient has similar problems and facial features as proband RB_078381. Learning disability is milder but did not do well at school. Short with a small head size on 3rd centile. | father of RB_078381 |
| RB_078381 | Manchester | deletion | 2 | Born at term weighing 3.16 kg. Poor growth; now short stature 0.4th – 2nd centile. Microcephalic with OFC below 3rd centile. Heart murmur investigated by echocardiogram and found to be innocent. Dysmorphic facial features; long columella, prominent incisors with wide gap between front teeth. Broadish thumbs and great toes, prominent fetal pads on fingers. Learning disability with IQ of 65 (mild MR), attends special school. Poor concentration. Normal echocardiogram. Exophoria. Normal neurologic examination. | |
| SAL_703574 | Salisbury | deletion | 2 | Referred ectrodactyly, ectodermal dysplasia, clefting syndrome; re-referred for array CGH with ectrodactyly on left hand, absent left foot. Mild global delay. Height, weight, and OFC all <0.4th percentile for age. Epicanthic folds, mild micrognathia, high palate, bifid uvula, turricephaly. Talipes of the right foot. Normal heart. Duane anomaly. Truncal hypotonia. Also have photos. | |

**Supplementary Table 7. Primers used in RACE and RT-PCR experiments.**

**a) RACE primers grouped by target region (GSP: gene-specific primer; 5' GSP refers to positive control primer)**

5' RACE

| target name | outer GSP | inner GSP | 5' GSP |
|---|---|---|---|
| 4.1 | TTTGTGCCTGCAGAAATGCCA | GCCTCTGGGGTAGAAGAAATGT | GCGATTAGAGCGGTTCAAACAA |
| 4.2 | ATGGCAGCTCCATAGAGAGA | ACAAACACCTTTTCACCTGTGT | ATGTGGGAGAGTCCATGCAA |
| 4.3 | AATGTGCCCCCAAATTGAGTTC | TGGAGCAAAGGGCTTATCTGAA | GACGATTGAACCCAATGAAGGC |
| 4.4 | GTGTTTTTGCCAGTCTCTGCTC | TGAGGGTGTAAGCATTCGGTTT | GCTGTTCAGATGGGATTGGTCA |

3' RACE

| target name | outer GSP | inner GSP |
|---|---|---|
| 6.1 | CTAGCCATAACTTGGTTGCATT | CTTGGTTGCATTCTCTAATCCG |
| 6.2 | GGCTTTGAGTTCAAGGTTCTGAC | AGTTCAAGGTTCTGACTGACCA |
| 6.3 | AACAGAGGCGAAACACAAATCC | CCTGGAGGAAAAGGCCTTGAA |
| 6.4 | CTTTGCCACGGTGTCCTTCA | CGCAGATCATGCAGAACTACCA |
| 6.5 | CCAAGAACCGGAAAGGCATC | GGCATCGCCATTATCATTCACG |
| 6.6 | TATAGGGCCTGTGATGATCTGA | AGAGGAGAAGGATGAGACTGATGA |
| 6.7 | TCTGGGAATGGAGGAAGAGACT | CCTCAAGCTGTCCTGCCTTC |
| 6.8 | GGTGACCTTCTCCATCATCGTG | GATAACCCAGCCTTCACCATTC |

**b) Primers used in nested targeted RT-PCR of *HYDIN2* transcripts**

Targeted RT-PCR

| | outer fwd | inner fwd | outer rev | inner rev |
|---|---|---|---|---|
| SMRT cell 1 | CATCCCTTCCAGCCTCAG | GGCCTTGGAAAATCGGAGC | CAGTGTGTCATTCAGGTTGG | TGGGTTCAATCGTCCAAAGG |
| | AGGACTACTGAGCAAGGC | GGCCTTGGAAAATCGGAGC | TTGAGTTCTGGAGCAAAGGG | TGGGTTCAATCGTCCAAAGG |
| SMRT cell 2 | CATCCCTTCCAGCCTCAG | GGCCTTGGAAAATCGGAGC | CTTCATGGCTGCGTAATCC | TCCGAGCAAAGAGAGTGTCG |
| | CCCTTTGGACGATTGAACC | AGCAGTACCTATCGGATTCC | CTTCATGGCTGCGTAATCC | TCCGAGCAAAGAGAGTGTCG |
| | CCAAGAACCGGAAAGGCATC | GGCATCGCCATTATCATTCACG | AGAAGACTCCTAGCTCATCC | CATTGTTCTTGAGGATGAGAGG |
| | GCAACGTGGGAAAGATCACC | TGTGCATTGCCAGTCATTCC | AGTTCCATCATCTTCCAGCC | TGTATAGGCCTGAGAGCTGC |
| | GCAACGTGGGAAAGATCACC | TGTGCATTGCCAGTCATTCC | AAGCTCCAGACATCCTCAGG | TCACCAGGAACCTTTGCC |

**c) Primers used for tissue expression measurements of *HYDIN* and *HYDIN2***

Expression RT-PCR

| | |
|---|---|
| GAPDH_fwd | TGAAGGTCGGAGTCAACGGATTTGGT |
| GAPDH_rev | CATGTGGGCCATGAGGTCCACCAC |
| ex46_fwd1 | CATGTACTGGGACCGGAAGC |
| ex46_rev2 | CCTTCAAAGTCTGGTGTCTGG |

**Supplementary Table 8. Fetal brain DNase I hypersensitivity samples with GEO accession numbers.**

| Sample ID | Tissue | Timepoint | Sex | Sequencing* | Project | Accession |
|---|---|---|---|---|---|---|
| DS16302 | fetal brain | d142 | F | se | Roadmap | GSM665819 |
| DS11877 | fetal brain | d122 | M | se | Roadmap | GSM595913 |
| DS11872 | fetal brain | d122 | M | se | Roadmap | GSM723021 |
| DS14464 | fetal brain | d117 | F | se | Roadmap | GSM595920 |
| DS15453 | fetal brain | d112 | U | se | Roadmap | GSM665804 |
| DS20231 | fetal brain | d109 | F | se | Roadmap | GSM878652 |
| DS23638 | fetal brain | d105 | M | pe | ENCODE | ENCBS493ZWQ |
| DS20780 | fetal brain | d105 | M | se | Roadmap | GSM1027328 |
| DS20226 | fetal brain | d104 | M | se | Roadmap | GSM878651 |
| DS20221 | fetal brain | d101 | M | se | Roadmap | GSM878650 |
| DS14803 | fetal brain | d96 | F | se | Roadmap | GSM595926 |
| DS14815 | fetal brain | d96 | F | se | Roadmap | GSM595928 |
| DS14717 | fetal brain | d85 | F | se | Roadmap | GSM595922 |
| DS14718 | fetal brain | d85 | F | se | Roadmap | GSM595923 |
| DS23849C | fetal brain | d80 | U | pe | ENCODE | ENCBS694XIX |
| DS24711A | fetal brain | d76 | M | pe | ENCODE | ENCBS980LUR |
| DS24872A | fetal brain | d72 | M | pe | ENCODE | ENCBS489VFT |
| DS24775A | fetal brain | d58 | M | pe | ENCODE | ENCBS852UJL |
| DS23813A** | fetal brain | d56 | U | pe | ENCODE | ENCBS539WGT |

*pe:paired end, se:single end

**shown in Figure 1e

**Supplementary Tables 9 and 10, which contain a list of MIP sequences used, can be found in additional file 3.**