

# Outcome of array CGH analysis for 255 subjects with intellectual disability and search for candidate genes using bioinformatics

Y. Qiao · C. Harvard · C. Tyson · X. Liu · C. Fawcett ·  
P. Pavlidis · J. J. A. Holden · M. E. S. Lewis ·  
E. Rajcan-Separovic

Received: 25 January 2010 / Accepted: 9 May 2010 / Published online: 29 May 2010  
© Springer-Verlag 2010

**Abstract** Array CGH enables the detection of pathogenic copy number variants (CNVs) in 5–15% of individuals with intellectual disability (ID), making it a promising tool for uncovering ID candidate genes. However, most CNVs encompass multiple genes, making it difficult to identify key disease gene(s) underlying ID etiology. Using array

CGH we identified 47 previously unreported unique CNVs in 45/255 probands. We prioritized ID candidate genes using five bioinformatic gene prioritization web tools. Gene priority lists were created by comparing integral genes from each CNV from our ID cohort with sets of training genes specific either to ID or randomly selected. Our findings suggest that different training sets alter gene prioritization only moderately; however, only the ID gene training set resulted in significant enrichment of genes with nervous system function (19%) in prioritized versus non-prioritized genes from the same de novo CNVs (7%,  $p < 0.05$ ). This enrichment further increased to 31% when the five web tools were used in concert and included genes within mitogen-activated protein kinase (MAPK) and neuroactive ligand-receptor interaction pathways. Gene prioritization web tools enrich for genes with relevant function in ID and more readily facilitate the selection of ID candidate genes for functional studies, particularly for large CNVs.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00439-010-0837-0) contains supplementary material, which is available to authorized users.

Y. Qiao · C. Harvard · E. Rajcan-Separovic (✉)  
Department of Pathology (Cytogenetics), Child and Family  
Research Institute, University of British Columbia (UBC),  
950 West 28th, Room 3060, Vancouver, BC V5Z 4H4, Canada  
e-mail: eseparovic@cw.bc.ca

Y. Qiao · M. E. S. Lewis  
Department of Medical Genetics, Child and Family Research  
Institute, UBC, Vancouver, BC, Canada

C. Tyson · C. Fawcett  
Cytogenetics Lab, Royal Columbian Hospital,  
New Westminster, BC, Canada

X. Liu · J. J. A. Holden  
Department of Psychiatry,  
Queen's University, Kingston, ON, Canada

P. Pavlidis  
Centre for High-Throughput Biology,  
UBC, Vancouver, BC, Canada

J. J. A. Holden  
Department of Physiology,  
Queen's University, Kingston, ON, Canada

J. J. A. Holden  
Cytogenetics and DNA Research Laboratory,  
Onganada, Kingston, ON, Canada

## Introduction

Intellectual disability (ID) (intelligence quotient  $< 70$ , with concordant deficits in adaptive function) comprises a highly heterogeneous group of disorders with an overall prevalence of ~1–3% (Chelly et al. 2006; Roeleveld et al. 1997). Genetic abnormalities are believed to be a major cause of ID. Up to 30% of cases were reported to be associated with chromosomal abnormalities (Knight et al. 1999) and hundreds of genes have been implicated (Kramer and van Bokhoven 2009). However, the etiology of ID still remains unknown in about half of all cases.

With the introduction of array-based comparative genomic hybridization (array CGH), which allows high-resolution whole genome analysis, an additional 5–15% ID patients

are found to carry pathogenic submicroscopic copy number variants (CNVs) (for review, see Koolen et al. 2009). With the increasing resolution and better understanding of CNV significance, array CGH is becoming an accepted tool in clinical genetics laboratories for finding the cause of ID. Compared to conventional cytogenetic and linkage analyses, array CGH greatly narrows down the potential pathogenic loci and is a promising tool for uncovering and cataloging ID candidate genes (Vissers et al. 2005). This is exemplified by the discovery of the *CHD7* gene in CHARGE syndrome (Vissers et al. 2004), mutations of which were identified after a sub-microscopic deletion was noted by array CGH. However, most of the CNVs encompass multiple (up to 100) genes, which makes it difficult to identify the key disease gene(s) for ID.

The availability of high throughput genome-wide datasets has facilitated and encouraged the rapid development of bioinformatics methods to prioritize positional candidate gene selection (Kanehisa and Bork 2003; Stein 2003). These methods are based on single or multiple data sources (see examples in Table 1). To minimize the bias from different single prioritization web tools, which also rely on different statistical methods, a combination of multiple web tools tends to be used for disease gene prediction (Elbers et al. 2007; Huang et al. 2008; Liu et al. 2008; Teber et al. 2009; Thornblad et al. 2007; Tiffin et al. 2006, 2008). Similarly, single gene prioritization web tools, which combine a large number of datasets, have been established. For example, the Endeavour prioritization web tool combines 26 datasets including gene ontology, literature, protein–protein interactions, sequence, gene expression datasets, etc. (Aerts et al. 2006). The prioritization studies, thus far have been mainly used for predicting and prioritizing the most promising candidate genes from genomic regions identified by linkage analyses and association studies in complex traits/diseases including type 2 diabetes and obesity (Elbers et al. 2007; Teber et al. 2009; Tiffin et al. 2006), osteoporosis (Huang et al. 2008; Liu et al. 2008) and metabolic syndromes (Tiffin et al. 2008). The application of web tools for gene prioritization in chromosomal regions identified by array CGH has only begun to be explored, with Endeavour being used as a single web tool (Osoegawa et al. 2008; Qiao et al. 2009; Yonan et al. 2003). More recently, the Ingenuity pathway analysis (IPA) web tool was used to help assess the candidate genes from the 1p34 array-detected microdeletion in a subject with autism (Kumar et al. 2009). Systematic analysis of the potential advantages and limitations of using different prioritization web tools to candidate gene identification in subjects with ID therefore remains largely unexplored.

In this study we used a selection of web tools to prioritize the genes within unique de novo, familial and common CNVs detected using array CGH in 255 subjects with ID.

**Table 1** Summary of Web site tools for candidate gene prioritization

Name	Website	Data source	Input	References
Endeavour	<a href="http://homes.esat.kuleuven.be/~bioiuser/endeavour/">http://homes.esat.kuleuven.be/~bioiuser/endeavour/</a>	Sequence, expression, functional annotation, protein interactions, pathways, literature mining, transcriptional motifs	Susceptibility loci or a list of test genes and a list of training genes	Aerts et al. (2006)
GeneWanderer	<a href="http://compbio.charite.de/genewanderer/">http://compbio.charite.de/genewanderer/</a>	Protein–protein interaction networks	Susceptibility loci and a list of training genes	Kohler et al. (2008)
PosMed	<a href="http://omicspace.riken.jp/PosMed/">http://omicspace.riken.jp/PosMed/</a>	Protein–protein interaction networks	Susceptibility loci and phenotype/disease keywords	Yoshida et al. (2009)
Suspects	<a href="http://www.genetics.med.ed.ac.uk/suspects/">http://www.genetics.med.ed.ac.uk/suspects/</a>	Literature mining	Susceptibility loci and phenotype/disease keywords	Adie et al. (2006)
ToppGene	<a href="http://toppgene.cchmc.org/">http://toppgene.cchmc.org/</a>	Mouse phenotype data, pathways, literature mining, protein interaction, expression	A list of test genes and a list of training genes	Chen et al. (2009)

Assessment of the function of the prioritized genes was performed using multiple databases including the mouse knockout phenotype database and other gene function analysis web tools. Such approaches are expected to help in the evaluation and understanding of the contribution of positional candidate genes involved in CNVs and their relationship to ID.

## Materials and methods

### Subjects

A total of 255 subjects with idiopathic ID were recruited for array CGH analysis by clinical geneticists across Canada. The criteria for selecting the cases included: (i) normal karyotypes by routine cytogenetic testing at the 500–550 band level resolution; (ii) negative fragile X testing by DNA analysis; (iii) a phenotype score  $\geq 3$  on a testing prioritization checklist adapted from de Vries et al. (2001); and (iv) both parents available for testing.

As much as 29 out of the 255 cases were reported as individual or small group cases (Gibson et al. 2008; Harvard et al. 2005; Rajcan-Separovic et al. 2007; Tyson et al. 2004,2005). Three array platforms were applied to screen the subjects: 1 Mb BAC array on 141 subjects, Agilent 105K Oligo array on 96 and NimbleGen array on 18 cases.

### Array CGH

Genomic DNA was extracted from peripheral blood using PUREGENE DNA Isolation Kit (Gentra, Minneapolis, MN). A pool of normal male or female control DNAs (Promega, Madison, WI) was used as reference DNA matching the sex of the proband samples.

#### *1 Mb resolution BAC array CGH*

BAC array CGH was performed as previously described (Rajcan-Separovic et al. 2007). Briefly, sample and reference DNAs were hybridized to the 1-Mb BAC array (Spectral Genomics, Houston, TX) using dye swap methods. Data analysis was performed using Spectralware 2 software (Spectral Genomics). Identification of clones with a significant gain or loss was based on previously established cutoff values of 1.2 and 0.8, respectively (Tyson et al. 2005).

#### *High-resolution oligonucleotide array CGH*

Agilent 105K array analysis was performed according to the protocol provided by the company (version 4.0, June 2006, Agilent Technologies, CA, USA) (Fan et al. 2007). Feature Extraction software (version 8.1.1.1, Agilent Tech-

nologies) rendered image analysis using the manufacturer's recommended settings (CGH\_-v4\_95) and human genome assembly hg18. The minimum absolute average of log<sub>2</sub> ratio was 0.25.

Higher-resolution 385K oligonucleotide genome array CGH was performed by courtesy of NimbleGen. Array log<sub>2</sub> ratio  $>\pm 0.2$  was used for a segmentation (region). For both the Agilent and NimbleGen array platforms,  $\geq 3$  consecutive probes were required for a significant CNV call. CNVs that overlapped in genomic coverage were considered to represent the same CNV loci.

### Criteria for interpreting CNVs

The criteria for the interpreting a CNV as unique or common have been described previously (Qiao et al. 2008). Briefly, CNVs overlapping with CNVs reported in at least two studies in the Database of Genomic Variants (DGV) or in our internal controls (Qiao et al. 2008) were considered common CNVs; those that overlapped partially (<50%) or did not overlap with CNVs reported in the DGV or our internal controls were called unique. Unique CNVs of de novo origin were considered pathogenic and unique CNVs of familial origin were considered putatively pathogenic. All unique CNVs were confirmed by FISH or custom array CGH and their parental origin determined using the same methods.

### Fluorescence in situ hybridization (FISH)

FISH analyses were performed using the BAC DNA clones from the CNVs, as described previously (Rajcan-Separovic et al. 2007). Slides were viewed on a Zeiss Axioplan 2 fluorescence microscope and images captured using MacProbe software (Applied Imaging, Santa Clara, CA).

### Custom oligonucleotide array CGH

For the validation of two maternally inherited abnormalities (duplications of 1p34.1 and 20q13.12), for which FISH testing was not possible, custom arrays were designed using eArray (Agilent technologies) and the ADM-2 algorithm as described previously (Rajcan-Separovic et al. 2010).

### Bioinformatics analysis

#### *Web tools for gene prioritization*

Five freely accessible gene prioritization web tools were selected: Endeavour (Aerts et al. 2006), GeneWanderer (Kohler et al. 2008), PosMed (Yoshida et al. 2009), Suspect (Adie et al. 2006) and ToppGene (Chen et al. 2009)

(Table 1). These five web tools use different data sources and require different inputs. Three of them (Endeavour, GeneWanderer and ToppGene) require a user-defined training-gene set (i.e. known disease genes), while two (PosMed and Suspects) automatically create their own “training”-gene set based on entering a phenotype specific term; in our case it was *mental retardation*.

#### Training-gene sets

For the three web tools that required user-defined training-gene sets, we selected six ID-related disease training-gene sets (ID 1–6) extracted from OMIM, Ensemble, Decipher, Suspects, GenTrepid, databases and an in-house ID training set which contained genes selected manually from ID-related publications (see details in Supplementary Table 1). In addition, six random training-gene sets (R1–R6) containing genes randomly selected from the whole human genome were used as negative controls (Supplementary Table 1). These genes were selected randomly using a random number generator from the “known genes” listed in the UCSC GoldenPath database (hg18) (<http://genome.ucsc.edu/>).

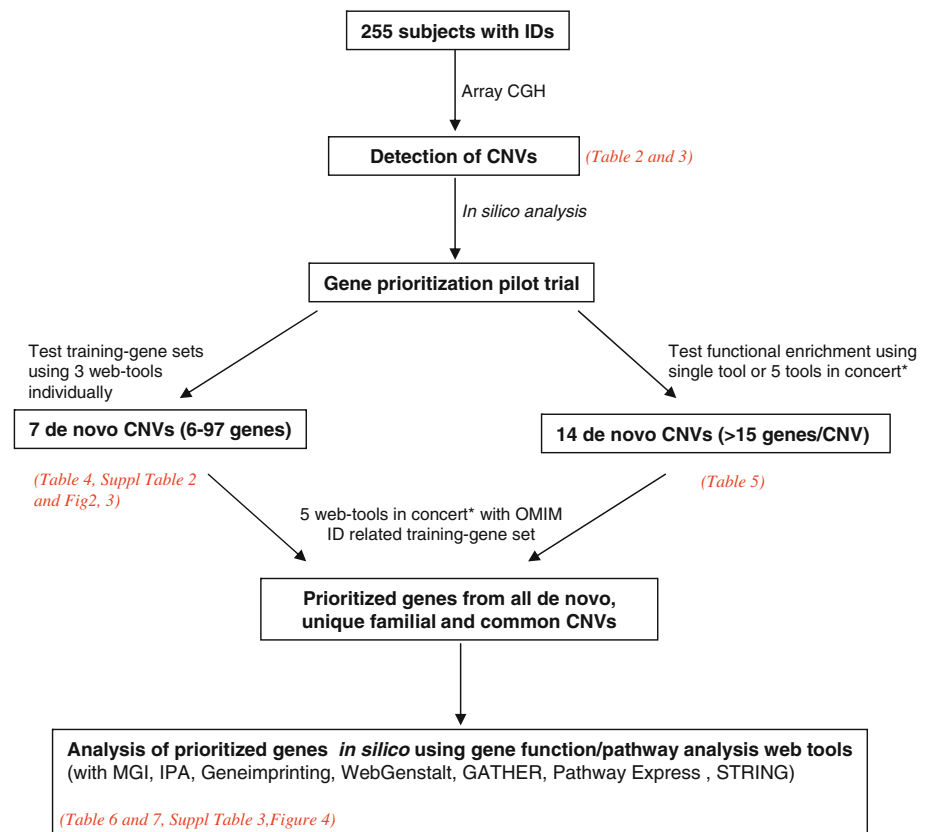
#### Overview of computational analysis (Fig. 1)

We first selected seven de novo CNVs (6–97 genes/CNV) for a pilot study to compare the outcome of gene prioritization using three web prioritization tools individually (Endeavour, GeneWanderer and ToppGene) with different training-gene sets. Next, we compared the function of prioritized and non-prioritized genes from 14 de novo CNVs (>15 genes/CNV); the genes were prioritized using a single web prioritization tool or five tools in concert. Finally, the function of the prioritized genes obtained using all five web tools in concert was compared between all de novo, familial and common CNVs.

#### Creation of priority lists and their analysis

(i) *7 pilot CNVs*: The priority list for a CNV consisted of the top five genes obtained using three web tools individually (Endeavour, ToppGene and GeneWanderer). As each individual web tool was used with six ID-specific and six random training sets separately, there were 12 web tool-specific priority lists for each tested CNV. In addition, as shown for one CNV from 9qter

**Fig. 1** Overview of the study



\* Indicates Endeavour, GeneWanderer, ToppGene, PosMed and Suspects

(Suppl Table 2), genes from the six priority lists obtained using six ID-specific training sets were ranked and the average of the rank for each gene was created. This “averaged ID priority list” is termed AP1 and contained, for each CNV, the top five genes with highest averaged rank. Similarly, for each CNV the genes from the six priority lists obtained using six random training sets were ranked and the average of the rank for each gene was used to create the averaged random priority list (AP2). Therefore, each CNV had 14 priority lists for each of the three tools: 12 individual gene priority lists obtained with six ID training sets and six random training sets, as well as AP1 and AP2. The corresponding priority lists were pooled for the seven CNVs per web tool (e.g. AP1 from CNV1 was pooled with AP1 from CNV2 and AP1 from CNV3, etc. for Endeavour). This created 14 web tool-specific pooled priority lists. These web tool-specific pooled priority lists were compared for gene overlap within each of the web tools (e.g. AP2 was compared with the remaining 13 priority lists for Endeavour). Also, the overlap of pooled priority list between any two web tools was determined (e.g. AP1 obtained with Endeavour was compared with AP1 obtained with Toppgene). Finally, overlap between AP1 lists and AP2 lists obtained by three tools was determined (e.g. AP1 obtained using Endeavour was compared to AP1 obtained using Toppgene and to AP1 obtained using Gene wanderer).

- (ii) *14 CNVs*: Five web tools in concert and one single web tool (Endeavour) were applied for prioritization of genes from 14 CNVs. For the “five tools in concert” approach the final priority list for a CNV consisted of genes which were among the top five in the priority lists obtained with at least two prioritization web tools (see an example for prioritization using five tools for CNV from 9qter in Supplementary Table 2). The pool of prioritized genes for 14 CNVs was compared in terms of gene function to non-prioritized genes from the same 14 CNVs using Ingenuity pathway analysis (IPA: <http://www.ingenuity.com>) software. IPA uses known protein–protein and gene–gene interactions in combination with multiple other data sources including differential expression from the microarray data.
- (iii) *All CNVs*: IPA software was used to compare the function of prioritized genes between three classes of CNVs as pools (de novo, familial and common) obtained using five web tools in concert with one training set (OMIM).

We also used three other publicly available web tools for pathway analysis of prioritized genes as a pool from all classes of CNVs (de novo, familial and common). These included;

- (1) WebGestalt (WEB-based GENE SeT AnaLysis Toolkit; <http://bioinfo.vanderbilt.edu/webgestalt/>) (Zhang et al. 2005);
- (2) Pathway-Express (<http://vortex.cs.wayne.edu/Projects.html>) (Draghici et al. 2003) and;
- (3) GATHER (<http://gather.genome.duke.edu/>) (Chang and Nevins 2006).

STRING (Search Tool for Retrieval of Interacting Proteins: <http://string.embl.de>) was used to identify and compare known and predicted protein–protein interactions based on comparative genomics and text mining (von Mering et al. 2007).

#### *Other public databases*

The Mouse Genome Informatics (MGI) resource (<http://www.informatics.jax.org>, version 3.54) (Bult et al. 2008; Eppig et al. 2007) was used to determine if human ID genes have mouse orthologs which show neurodevelopmental phenotypes resulting from the homozygous and/or hemizygous disruption of the gene(s) in knockout mouse models. The mouse phenotypes reviewed for prospective roles in ID etiology involved those affecting the nervous system, embryogenesis, as well as behavioral, neurological and craniofacial phenotypes.

The GeneImprinting website database (<http://www.geneimprint.org/site/genes-by-name>) was used to determine if genes, prioritized from familial inherited CNVs may be imprinted. Selective parental, monoallelic expression could explain the presence of the abnormal phenotype in an affected child that is absent in the transmitting parent (Lee et al. 2007).

#### Statistical analysis

All statistical tests were performed using the statistical computation Web site from Vassar College (<http://faculty.vassar.edu/lowry/VassarStats.html>). The chi-square and Fisher’s exact tests were used.

## Results

### Array CGH analysis

We identified 47 unique and previously unreported CNVs in 45 out of 255 subjects with ID (17.6%) tested using the three different array platforms (Table 2). Each CNV was confirmed either by FISH or by custom array analysis followed by parental analysis using the same methods: 21 CNVs were de novo (in 20 subjects or 7.8%) and 26 were familial CNVs (in 25 subjects or 9.8%) (Table 2). The

**Table 2** Summary of array CGH findings

	De novo CNVs	Familial CNVs	Total subjects
CNV prevalence in subjects			
Spectral genomics BAC (1 Mb)	6.40%	0.70%	141
Agilent 105K Oligo (0.02 Mb)	9.40%	20%	96
NimbleGene 385K Oligo (0.01 Mb)	11.10%	28%	18
Total	7.8%	9.8%	255
CNV characteristics			De novo + familial CNVs
No. of CNVs	21	26	47
Average size (Mb)	3.1	0.6	1.7
No. of non-redundant genes	595	116	711
Average No. of genes/CNVs	30	5	16
Ratio of del:dup	1.6:1	0.1:1	0.5:1

detection rate of de novo CNVs was higher (9.4 and 11.1%) with the higher-resolution arrays (Agilent 105K and Nimblegen 385K) compared to the low-resolution array CGH (6.4%, SG). Deletions were more frequent for de novo (13/21) than familial (1/26) CNVs ( $p < 0.0001$ ). Six de novo CNVs were found to overlap with known genetic microdeletion/duplication syndromes (Table 3). The CNV sizes range from 680 kb to 9.7 Mb for de novo (average 3.0 Mb) and 31 kb to 1.7 Mb for familial CNVs (average 0.6 Mb). The de novo CNVs contained 1–97 genes/CNV, while the familial CNVs contained 0–21 genes. In total 595 and 116 non-redundant genes were found to be involved in the 21 de novo and 26 familial CNVs, respectively (Table 3). The 17 most frequently recurring common CNVs (detected in >10% of ID subjects on each of the three array platforms (Qiao et al. 2008)) and overlapping with CNVs from the database of genomic variants (<http://projects.tcag.ca/variation/>) are also listed in Table 3. They contained a total of 108 genes.

#### Candidate gene prioritization in a selection of seven pilot CNVs—impact of training set and gene prioritization web tool on ID candidate gene priority lists

In our preliminary prioritization experiment, we selected seven de novo CNVs of various sizes (0.5–7.9 Mb) and number of integral genes (between 6 and 97) (marked in Table 3) to test the effect of training-gene sets on priority lists obtained with each of the three web tools (Endeavour, GeneWanderer and ToppGene) that require user-defined training sets. Overlap between AP2 (averaged random priority list) pooled for seven CNVs with pooled priority lists obtained using ID training set 1–6 ranged from 46 to 77% for the three tools used (Fig. 2a). It was even higher when

the individual priority lists, obtained using six random training sets, each pooled for the seven CNVs, were compared with AP2 lists pooled for seven CNVs (83–97%). The rate of overlap between the priority lists within the web tool depended on the number of genes/CNV as CNVs with more genes (>15 genes/CNV) showed less overlap between the averaged priority lists AP1 and AP2 (65–75%) compared to smaller CNVs with <15 genes/CNV (87–100%, Table 4). The >50% overlap between the priority lists obtained using random and ID training-gene sets for the seven CNVs suggests limited disease specificity for the tools regardless of the training-gene set.

The results of the prioritization were further investigated by assessing overlap between prioritized genes in AP1 and AP2 lists (each pooled for seven CNVs) among the three tools. A low number of genes in common was noted when any two of the three tools were compared and ranged between 26 and 45% (Fig. 2b). Similarly, there were few genes common to all three tools within pooled AP1 and AP2 list (5/76 and 6/72) (Fig. 3). This indicated a significant discrepancy in prioritization results among the different tools. The majority of these common genes (4) was the same between AP1 and AP2 priority lists (Fig. 3). Most of these common genes (3/4) were well-investigated disease-related genes (*CRHR1* (Varela et al. 2006), *MAPT* (Koolen et al. 2008), *MBD5* (Jaillard et al. 2009)), suggesting that both the random and the ID-specific training sets tend to prioritize known disease genes.

#### Candidate gene prioritization in a selection of 14 CNVs—comparison of function of prioritized and non-prioritized genes

Based on the observation that only 26–45% of genes overlap for any two tools that require training sets (Fig. 2b), we prioritized genes from 14 larger CNVs (>15 genes) using five tools in concert and compared them with gene priority lists obtained with one individual tool as well as with non-prioritized genes. We selected Endeavour as a single tool, because it captures the largest number of data sources and utilizes more flexible input forms with validated in vivo experiments on a prioritized gene (Tranchevent et al. 2008). Prioritized gene lists were obtained using only one of the ID-specific training sets (OMIM) and one random training set (R1). The OMIM training set was selected because it can be automatically downloaded from the frequently cited and updated Web site with the option of selecting genes with known sequence and phenotypes (Supplementary Table 1).

The outcome of the prioritization was evaluated by analysis of the function of the prioritized genes using IPA software. The gene function comparison was made between the initial non-prioritized 491 genes derived from the 14 CNV loci to pooled prioritized genes from these CNVs

**Table 3** List of de novo, familial and most common CNVs and prioritized genes

CNV group	No.	Cytoband and CNV	Subjects	Start (bp)	End (bp)	Size (bp)	No. of genes	The top five prioritized genes
A. de novo	1	1p36 dup	06-01 <sup>AG</sup>	26,575,529	27,392,834	817,305	16	ARID1A <sup>b</sup> , FAM46B, NR0B2 <sup>a</sup> , RPS6KA1 <sup>a</sup> , SLC9A1 <sup>a,b</sup>
	2	2p15.1-16 del <sup>c,d</sup>	03-02 <sup>SG</sup> & 03-57 <sup>SG</sup>	55,499,483	63,368,196	7,868,713	26	EFEMP1, OTX1 <sup>a,b</sup> , PEX13 <sup>b</sup> , PNPT1, REL <sup>b</sup> , XPO1
	3	2p13 del	06-108 <sup>AG</sup>	72,140,702	72,924,626	783,924	2	CYP26B1 <sup>a,b</sup> , EXOC6B <sup>a</sup>
	4	2q23.1 del <sup>c,d</sup>	07-21 <sup>AG</sup>	148,595,892	150,837,093	2,241,201	6	C2orf25, EPC2, KIF5C <sup>a,b</sup> , LYPD6, LYPD6B, MBD5 <sup>a,b</sup>
	5	5q14.1 dup	05-05 <sup>NG</sup>	80,370,000	90,150,000	9,780,000	24	ACOT12, CCNH <sup>a</sup> , HAPLN1 <sup>a,b</sup> , MEF2C <sup>a,b</sup> , RASA1 <sup>a,b</sup> , VCAN <sup>a</sup> , XRCC4 <sup>a,b</sup>
	6	7q36.3-7qter del <sup>d</sup>	02-06 <sup>SG</sup>	154,930,000	158,600,000	3,670,000	19	EN2 <sup>a,b</sup> , MNX1 <sup>a,b</sup> , PTPRN2 <sup>a</sup> , SHH <sup>a,b</sup> , UBE3C, VIPR2 <sup>a,b</sup>
	7	9q21.1 dup	05-49 <sup>NG</sup>	78,630,000	81,090,000	2,460,000	7	GNAI4 <sup>a</sup> , GNAQ <sup>a,b</sup> , PSATI, VPS13A <sup>a,b</sup>
	8	9qter del <sup>c,d</sup>	06-MDL-3 <sup>SG</sup>	139,523,178	140,007,383	484,205	9	CACNA1B <sup>a,b</sup> , EHMT1 <sup>a,b</sup> , PNPLA7, WDR85, ZMYND19 <sup>a</sup>
	9	10p12.2 del <sup>e</sup>	07-27 <sup>AG</sup>	26,714,221	29,234,923	2,520,702	16	ABI1 <sup>a</sup> , ACBD5, ANKRD26 <sup>a,b</sup> , BAMBI <sup>a</sup> , PDSS1, YMEIL1 <sup>a</sup>
	10	10q21.1 del	07-21 <sup>AG</sup>	57,249,135	58,942,307	1,693,172	1	ZWINT <sup>a</sup>
	11	11q12.3-13.1 dup	02-27 <sup>SG</sup>	62,270,860	64,896,754	2,625,894	90	PRDX5 <sup>a</sup> , PYGM, SF1 <sup>a</sup>
	12	11q24 del	05-40 <sup>SG</sup>	124,290,000	129,030,000	4,740,000	37	CHEK1 <sup>a,b</sup> , ETS1, KCNJ5 <sup>a</sup> , RICS <sup>a,b</sup> , ST3GAL4
	13	11q24 dup	04-48 <sup>SG</sup>	123,150,000	128,190,000	5,040,000	62	CHEK1 <sup>a,b</sup> , NRG1 <sup>a,b</sup> , ST3GAL4
	14	11qter del	04-48 <sup>SG</sup>	128,250,000	134,425,035	6,175,035	32	ACAD8 <sup>a</sup> , ADAMTS8 <sup>a</sup> , APLP2 <sup>a,b</sup> , B3GAT1 <sup>a,b</sup> , JAMS3 <sup>a</sup> , OPCML <sup>a</sup>
	15	16p13.3 del <sup>d</sup>	04-36 <sup>SG</sup>	1	1,094,450	1,094,449	51	AXIN1 <sup>a,b</sup> , HBA1 <sup>b</sup> , MPG, SOX8 <sup>a</sup> , STUB1 <sup>a,b</sup>
	16	16q22 del	06-14 <sup>AG</sup>	69,620,645	73,055,007	3,434,362	24	ATBF1 <sup>a</sup> , CHST4, DHODH, DHX38, HPR, KIAA0174, TAT
	17	17q12 del	06-99 <sup>AG</sup>	31,503,207	33,323,172	1,819,965	24	AATF <sup>a,b</sup> , ACACA <sup>a,b</sup> , DDX52, LHX1 <sup>a,b</sup> , TADA2L, TCF2 <sup>a,b</sup>
	18	17q21 del <sup>c,d</sup>	06-88 <sup>AG</sup>	41,011,330	41,700,962	689,632	8	CRHR1 <sup>a,b</sup> , IMP5, KIAA1267 <sup>a</sup> , MAPI <sup>a,b</sup> , RPS26P8
	19	17q25.3 dup	08-38 <sup>AG</sup>	77,660,313	78,154,619	494,306	12	CI7orf101, CD7 <sup>b</sup> , CSNK1D <sup>a,b</sup> , FOXK2, HEXDC, SLC16A3, UTS2R <sup>a</sup>
	20	19p13 dup <sup>c</sup>	05-37 <sup>SG</sup>	1,709,657	5,057,195	3,347,538	97	AES <sup>a,b</sup> , EEF2 <sup>a</sup> , OAZ1, SIPR4 <sup>a,b</sup> , SIRT6, THOPI <sup>a</sup>
21	20q13.33 dup <sup>c</sup>	06-139 <sup>AG</sup>	58,699,464	62,363,774	3,664,310	68	CHRNA4 <sup>a,b</sup> , LAMA5 <sup>a,b</sup> , OPR1 <sup>a,b</sup>	
B. Familial	1	1q21.1 dup (mat) <sup>d</sup>	04-18 <sup>NG</sup>	145,110,000	146,190,000	1,080,000	10	ACP6, BCL9, CHD1L, FMO5, GJA5 <sup>a</sup> , GJA8 <sup>a</sup> , GPR89B, PRKAB2
	2	1q21 dup (pat) <sup>d</sup>	08-22 <sup>AG</sup>	144,510,700	146,294,854	1,784,154	14	ACP6, FMO5, GJA5 <sup>a</sup> , GJA8 <sup>a</sup> , PRKAB2
	3	1p34.1 dup (mat)	07-53 <sup>AG</sup>	45,014,474	45,200,817	186,343	9	EIF2B3, PLK3 <sup>a</sup> , PTCH2 <sup>a,b</sup> , RPS8
	4	1q44 dup (pat)	06-67 <sup>NG</sup>	244,230,000	244,530,000	300,000	1	SMYD3
	5	2p12 dup (mat)	06-40 <sup>AG</sup>	78,428,495	79,496,295	1,067,800	4	REG1A, REG1B <sup>b</sup> , REG3A <sup>a</sup> , REG3G
	6	3p26.2 del (pat)	07-32 <sup>AG</sup>	4,206,159	4,434,847	228,688	2	SETMAR, SUMF1 <sup>a,b</sup>
	7	4q35 dup (mat)	05-01 <sup>AG</sup>	187,178,973	187,370,217	191,244	4	TLR3, CYP4V2, FAM149A, FLJ138576 <sup>a</sup>
	8	4q32 dup (mat)	05-MDL-8 <sup>AG</sup>	160,736,514	162,776,644	2,040,130	1	FSTL5
	9	5p12 del (pat),	04-14 <sup>AG</sup>	12,517,732	12,802,101	284,369	0	No gene
	10	7q31.1 dup (pat)	07-01 <sup>AG</sup>	117,187,756	118,908,220	1,720,464	3	ANKRD7, CTNBP2 <sup>a</sup> , LSM8
	11	7q31.1 dup (mat)	04-38 <sup>NG</sup>	113,010,000	113,430,000	420,000	1	PPP1R3A

Table 3 continued

CNV group	No.	Cytoband and CNV	Subjects	Start (bp)	End (bp)	Size (bp)	No. of genes	The top five prioritized genes
	12	10q22 dup (mat)	08-40 <sup>AG</sup>	74,661,716	74,809,824	148,108	4	MRPS16 <sup>a</sup> , ANXA7 <sup>b</sup> , TTC18, KIAA0974
	13	10q dup (pat)	05-36 <sup>NG</sup>	76,353,000	77,439,000	1,086,000	10	COMTD1, DUPDI, DUSP13, MYST4 <sup>a, b</sup> , SAMD8, VDACC2 <sup>a</sup> , ZNF503 <sup>a</sup>
	14	11q dup (pat)	03-24 <sup>NG</sup>	133,845,000	134,449,035	604,035	1	LOC729305
	15	12p13 dup (mat)	03-MDL-24 <sup>AG</sup>	5,952,451	6,280,226	327,775	2	CD9 <sup>a</sup> , VWF
	16	14q23 del (pat)	05-MDL-7 <sup>AG</sup>	72,245,680	72,400,074	154,394	1	DPF3
	17	15q21 dup (mat)	06-54 <sup>AG</sup>	42,843,447	43,520,794	677,347	13	DUOX1, DUOX2 <sup>b</sup> , GATM <sup>a</sup> , SLC28A2 <sup>a</sup> , SORD <sup>b</sup>
	18	16p11.2 dup (pat) <sup>d</sup>	06-32 <sup>AG</sup>	29,500,084	30,027,413	527,329	25	ALDOA, CDIPT <sup>a</sup> , MAZ <sup>a</sup> , PPP4C <sup>b</sup> , QPRT <sup>a</sup>
	19	17p dup (mat)	04-14 <sup>AG</sup>	10,557,304	11,130,994	573,690	3	TMEM220
	20	18p11 dup (mat),	06-40 <sup>AG</sup>	9,985,330	10,890,656	905,326	5	APCDD1 <sup>a</sup> , C18orf30, C18orf58, FAM38B, NAPG
	21	20q dup (mat)	06-01 <sup>AG</sup>	41,768,223	41,798,860	30,637	2	FAM112A, MYBL2 <sup>b</sup>
	22	21q22 dup (pat)	06-117 <sup>AG</sup>	34,648,298	34,821,005	172,707	4	C21orf51, DSCR1 <sup>a</sup> , KCNE1 <sup>a,b</sup> , KCNE2 <sup>a</sup>
	23	Xq12 dup (pat)	07-29 <sup>AG</sup>	65,732,015	66,158,034	426,019	1	EDA2R
	24	Xp22.31 dup (mat)	03-25 <sup>SG</sup>	7,565,092	8,091,951	526,859	2	PNPLA4, VCX
	25	Xq22 dup (mat)	07-99 <sup>AG</sup> & 07-96 <sup>AG</sup>	103,072,582	103,190,187	117,605	2	H2BFM, H2BFWT
C. Most common	26	Xp11.2 dup (mat)	07-91 <sup>AG</sup>	50,842,514	51,096,143	253,629	1	NUDT10
	1	1p21.1	SG	103,956,050	104,113,271	157,221	4	AMY1A, AMY1B, AMY1C, AMY2A
	2	1q31.3	AG	195,011,144	195,104,407	93,263	2	CFHR1, CFHR3
	3	1q44	AG	246,442,498	246,806,773	364,275	14	OR2G6, OR2M4, OR2M7, OR2T1, OR2T12, OR2T2, OR2T29, OR2T3, OR2T33, OR2T34, OR2T4, OR2T5, OR2T6
	4	2p11.2	NG	89,370,000	89,910,000	540,000	2	IGKC, IGKV1D-42
	5	3q26.1	NG, AG	163,940,972	164,109,000	168,028	0	No gene
	6	4q13.2	NG, AG	68,901,010	69,166,014	265,004	3	TMPRSS11B, UGT2B17
	7	6p21.32	AG	32,519,735	32,673,183	153,448	4	HLA-DRB1, HLA-DRB5, HLA-DRB9
	8	8p23.1	AG	7,261,218	8,132,339	871,121	25	DEFB103A, DEFB104B, DEFB106A
	9	8p11.23	AG	39,341,324	39,499,952	158,628	1	ADAM5P
	10	9q32	SG	114,865,612	114,972,775	107,163	2	SLC31A2, FKBP15
	11	10q11.22	NG, AG	45,455,157	47,735,672	2,280,515	17	ANXA8, FAM21B, PPYR1 <sup>a,b</sup>
	12	13q21.1	SG	56,650,362	56,729,961	79,599	0	No gene
	13	14q11.1	NG, AG	18,090,000	19,497,223	1,407,223	11	OR11H12, OR4K2, OR4K5, OR4M1, OR4N2
	14	15q11.2	NG, AG	18,261,000	20,317,192	2,056,192	10	HERC2P3, OR11K1P, POTE8, Q8N9W7
	15	16p11.2	NG, AG	31,804,684	33,758,282	1,953,598	7	SLC6A10P <sup>a</sup> , ZNF267



Table 3 continued

CNV group	No.	Cytoband and CNV	Subjects	Start (bp)	End (bp)	Size (bp)	No. of genes	The top five prioritized genes
	16	17q21.31	NG, AG	41,550,000	41,730,000	180,000	4	KIAA1267 <sup>a</sup> , ARL17P1, LRRC37A2
	17	19p13.2	SG	8,743,041	8,864,039	120,998	2	ZNF558, MBD3L1

AG detected by Agilent 105K array platform, NG detected by NimbleGen 385K array platform, SG detected by Spectral Genomics 1 Mb BAC array platform, Mat maternally inherited, Pat paternally inherited

<sup>a</sup> PubMed manual search showed brain-related function

<sup>b</sup> MGI showed nervous system phenotype in knockout mice

<sup>c</sup> 7 CNV loci used in pilot prioritization experiment

<sup>d</sup> CNV overlapped with known genetic microdeletions/duplications syndromes

obtained with the above described approaches (Endeavour vs. five gene prioritization web tools in concert; OMIM vs. random set). Our results show that genes in the IPA category with biological function were significantly more prevalent after prioritization using Endeavour with either OMIM disease training-gene set (89%) or R1 random training-gene set (78%) compared to non-prioritized genes (53% had biological function) (both  $p < 0.001$ ; Table 5). Similarly, enrichment for genes in the IPA category “with nervous system development and function” was noted both with OMIM and with random training set (19 and 13%, respectively) compared to non-prioritized genes (7%); however, it was significant only after prioritization using Endeavour with the OMIM training-gene set ( $p = 0.004$ ), and not with random training-gene set R1 ( $p = 0.1$ ).

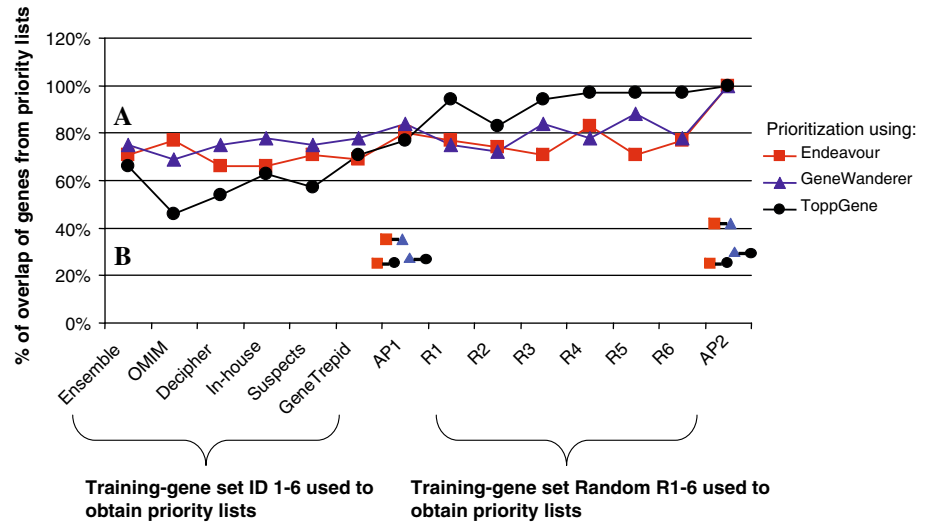
When the results of prioritization using five tools were combined as described previously (with OMIM disease training-gene set), the prevalence of genes with biological/nervous system function after prioritization was even higher, as 93 and 31% of prioritized genes had biological and nervous system functions, respectively, in comparison to 53 and 7% of non-prioritized genes ( $p < 0.0001$ ; Table 5).

Candidate gene prioritization from all de novo, familial and most common CNVs including analysis of the function of prioritized genes

Based on these results, we used the same five tools in combination method and the OMIM disease training-gene set to prioritize the genes in each of our remaining CNVs. For CNVs with  $\leq 5$  genes, we “prioritized” all of them as ID candidate genes. In summary, 102 out of 595, 71 out of 116 and 51 out of 108 genes represented candidate ID genes from 21 de novo, 27 familial and 17 most common CNVs after prioritization (Tables 3, 6).

We used a number of databases and bioinformatics web tools to compare the function of genes prioritized from the different CNV groups. First, we applied the MGI database to determine if loss of function of any of the prioritized genes (due to either homozygous or hemizygous disruption in mouse models) caused mouse phenotypes. We found that 37 out of 66 (56%) prioritized genes from 13 de novo deletion CNVs had been investigated in knockout mouse model on MGI. Among them, 70% (26/37) were found to have annotations related to mouse knockout phenotypes including nervous system abnormalities, abnormalities during embryogenesis, behavior/neurological phenotypes, and/or craniofacial phenotypes (Table 3), suggesting an enrichment of genes with these functions in the priority list obtained for the de novo deletion CNVs. It is interesting that only ~25% of all human genes currently have mouse ortholog knockout data and 70% of our priority genes were

**Fig. 2** Overlap of gene priority lists. **A** The percentage of overlap presented for each web tool separately and obtained by comparing the averaged random priority lists (AP2) pooled for seven CNVs to each of the 13 priority lists pooled for seven CNVs. These 13 lists were based on using ID training sets 1–6, random training sets 1–6 and the averaged ID priority list (AP1). **B** Overlap of averaged ID priority list (AP1) or averaged random priority list (AP2) (pooled for seven CNVs) between any two out of the three web tools

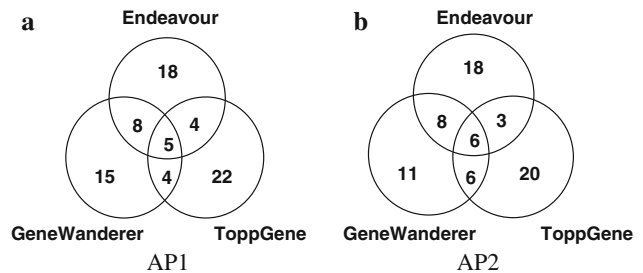


**Table 4** Rate of overlap between the averaged priority lists (AP1 and AP2) from seven de novo CNVs

	Endeavour	GeneWanderer	ToppGene
For all loci (7 CNVs)	28/35 (80%)	27/32 (84%)	27/35 (77%)
For loci with >15 genes (4 CNVs)	13/20 (65%)	15/20 (75%)	14/20 (70%)
For loci with <15 genes (3 CNVs)	15/15 (100%)	12/12 (100%)	13/15 (87%)

among them. We could not assess the presence of prioritized genes from our familial and common CNVs in MGI phenotypes as were predominantly involved in duplications.

Next, we used IPA software to compare the enrichment of genes prioritized from different CNV groups. We found a significantly higher proportion of genes with biological and nervous system function in the de novo CNV subgroup compared to the familial or most common CNVs (Table 6;  $P < 0.05$ ). The genes from familial CNVs identified to have a role in nervous system function were *KCNE1* (Letts et al. 2000), *CD9* (Doh-ura et al. 2000) and *REG1B* (Tebar et al. 2008). We also searched for evidence of imprinting for familial prioritized genes based on information from the



**Fig. 3** The number of prioritized genes overlapping between three web tools. **a** The number of genes from AP1 lists pooled for seven CNVs detected with each of the three web tools and their overlaps. The five genes detected by all web tools are: *CRHR1*, *EHMT1*, *EPC2*, *MAPT* and *MBD5*. **b** The number of genes from AP2 lists pooled for seven CNVs detected with each of the three webtools and their overlaps. The six genes detected by all web tools are: *ACBD5*, *CRHR1*, *EPC2*, *MAPT*, *MBD5* and *MPP7*

*Geneimprinting* database (<http://www.geneimprint.org/site/genes-by-name>). Only one gene, *GATM*, showed imprinting in mouse that is of unknown imprinting status in humans. By searching gene ontology terms, four genes (*DPF3*, *MYST4*, *SETMAR* and *SMYD3*) were found to have chromatin modification functions in the familial CNV group, while three genes (*ARID1A*, *EHMT1* and *EPC2*) with chromatin modification function were found in the

**Table 5** IPA analysis of the function of prioritized genes from a selection of 14 de novo CNVs

	Non-prioritized	Prioritized		
		Endeavour/OMIM disease set	Five tools/OMIM disease set	Endeavour/random set
No. of genes	491	64	71	64
Genes with biological function	53% <sup>abc</sup>	89% <sup>a</sup>	93% <sup>cd</sup>	78% <sup>bd</sup>
Genes with NS function	7% <sup>ef</sup>	19% <sup>e</sup>	31% <sup>fg</sup>	13% <sup>g</sup>

NS nervous system

<sup>a-g</sup> Indicates the pair of comparisons showing a significant difference ( $p < 0.05$ )

**Table 6** IPA analysis of the function of prioritized genes from different CNV group

	De novo CNVs	Familial CNVs	Most common CNVs	OMIM training genes
Total prioritized gene number	102	71	51	61
Genes with biological functions	87%	59%	29%	93%
Genes with NS function	24%	6%	0	32%

All the differences between de novo and familial CNVs showed statistical significance ( $p < 0.05$ )

NS nervous system

**Table 7** The two pathways enriched for genes prioritized from our de novo CNVs

Pathway	Gene	Cytoband
MAPK signaling pathway	RPS6KA1 <sup>a</sup>	1p36
	RASA1 <sup>b</sup>	5q14 <sup>c</sup>
	MEF2C <sup>a, b</sup>	5q14 <sup>c</sup>
	CACNA1B <sup>a, b</sup>	9qter <sup>c</sup>
	MAPT <sup>a, b</sup>	17q21.1 <sup>c</sup>
Neuroactive ligand–receptor interaction	VIPR2 <sup>b</sup>	7q36.3 <sup>c</sup>
	CRHR1 <sup>a, b</sup>	17q12–q22 <sup>c</sup>
	UTS2R	17q25.3
	S1PR4 <sup>b</sup>	19p13.3
	OPRL1 <sup>b</sup>	20q13.33

<sup>a</sup> Indicates the genes reported in published ID studies

<sup>b</sup> Indicates the genes with comparable nervous system phenotype in knockout mice in MGI

<sup>c</sup> Indicates deletion CNV

novo CNV group. No gene with imprinting or chromatin-related functions was found in the prioritized group from the most common CNVs.

To analyze the pathways involving the prioritized genes, we applied IPA software and three public web-pathway tools (WebGenstalt, Pathway-Express and Gather) (Supplementary Table 3). The significant enrichment of prioritized genes from de novo CNVs in neuroactive ligand-receptor interaction and MAPK signaling pathways was detected by two pathway tools (Webgestalt and Pathway express). The prioritized genes from our de novo CNVs involved in these two pathways are listed in Table 7. We also tested the six ID training-gene sets for participation in pathways and, in addition to the two pathways listed above, enrichment was detected in regulation of actin cytoskeleton development and axonal guidance by more than one web tool. No enrichment for any specific pathway was found among the random training-gene sets. Genes in the familial and most common CNVs were found to be more involved in pathways related to carbohydrate metabolism and immune responses (Supplementary Table 3).

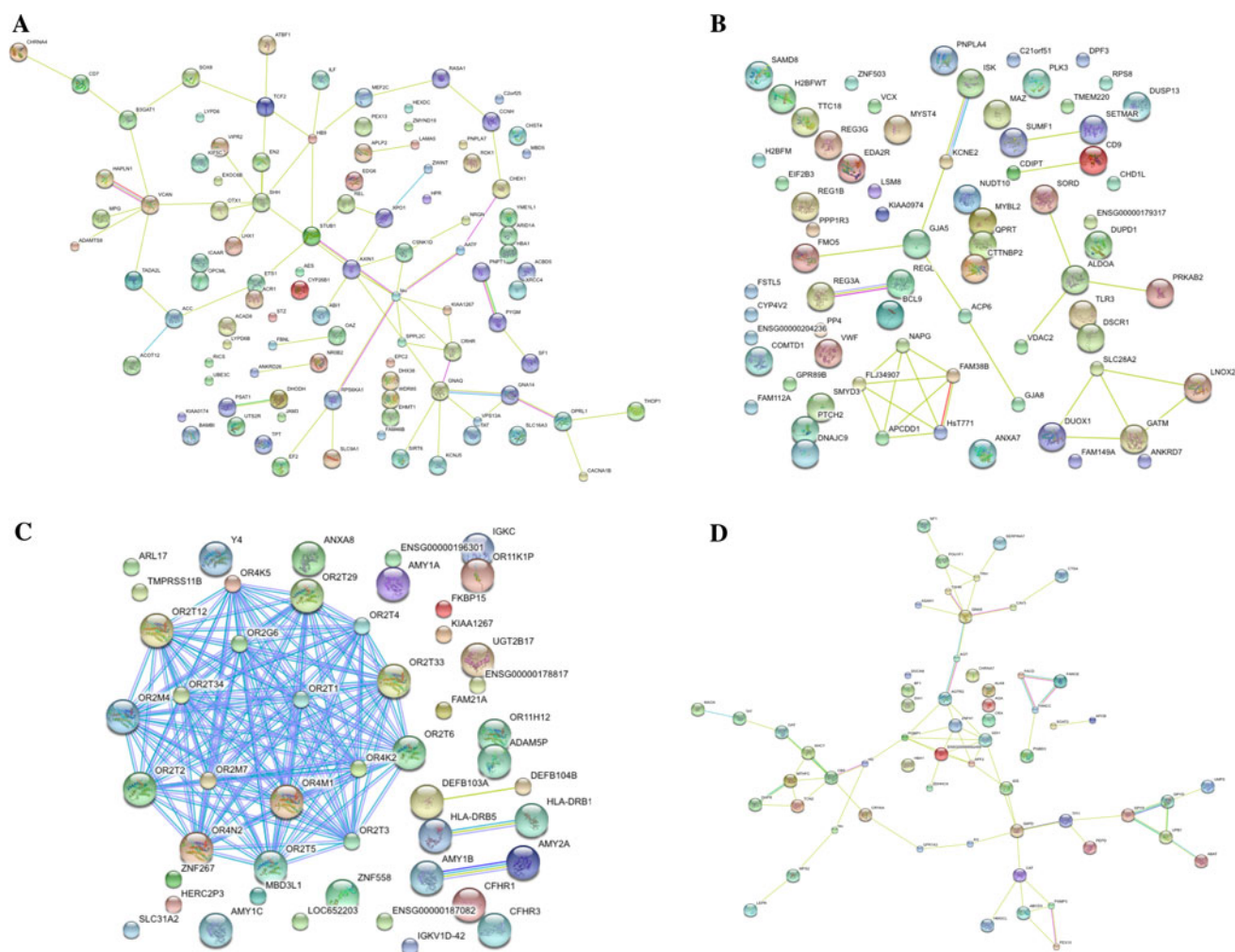
Finally, we analyzed the network of interactions of prioritized genes from each of the three CNV groups and the OMIM training set using STRING (Fig. 4). We noted that

the genes from the OMIM training set and de novo CNVs are more likely to have connections compared to familial and most common CNVs. There were 8/61 (13%) isolated genes (i.e. without any connection to other genes) in the OMIM set, 33/102 (32%) in the de novo, 44/71 (62%) in the familial CNVs and 26/51 (51%) in the common CNV set.

## Discussion

We used three array CGH platforms to screen for unique CNVs in 255 subjects with ID. Our total detection rate of de novo CNVs (7.8%) is consistent with previously reported results (5.7–11%) (Koolen et al. 2009), and increased to 9.6% as the resolution of array platforms increased. As expected, the higher-resolution arrays detected smaller CNVs (the smallest being 31 kb). These small CNVs appear to be mostly familial in origin (for CNVs < 500 kb in size, 15/27 were from familial CNVs vs. 2/21 from de novo CNVs, Table 3), while the de novo CNVs tended to be larger (on average, 3.1 Mb for de novo vs. 0.6 Mb for familial CNVs) and contain a larger number of genes (30 genes per de novo CNV vs. 5 genes per familial CNV, on average, Table 2). Two exceptions are the de novo deletions of 2p13 (780 Kb) and 10q (1.6 Mb), which harbored two (*CYP26B1* and *EXOC6B*) and one (*ZWINT*) gene, respectively. Disruption of *CYP26B1* has been reported to affect neural crest formation and the knockout mouse showed abnormal craniofacial morphology (Maclean et al. 2009). *EXOC6B* is involved in brain exocytosis (Brymora et al. 2001), and *ZWINT* is highly expressed in the brain, localized extensively in primary hippocampal neurons (van Vlijmen et al. 2008). The remaining de novo CNVs contain at least six genes/CNV.

By using five gene prioritization tools, we prioritized the ID candidate genes from each unique de novo, familial and most common CNVs. We identified the candidate ID genes from each CNV separately in order to avoid the over-representation of genes from larger chromosomal regions with more genes and the under-representation of genes from small CNV loci with few genes, which could occur if prioritization of all genes is performed as a pool.



**Fig. 4** Interactions of prioritized genes in different CNV groups. **a** Prioritized genes from de novo CNVs. **b** Prioritized genes from familial CNVs. **c** Prioritized genes from most common CNVs. **d** All genes from OMIM training set

One of the challenges in using prioritization tools is the discrepancy in candidate gene selection resulting from using different web tools (Teber et al. 2009; Thornblad et al. 2007). For example, two studies by Tiffin et al. (2006) and Teber et al. (2009) used the same initial group of unprioritized 9,556 positional candidate genes and slightly different web tools (six out of eight web tools were the same for the two studies) to predict candidate genes for type 2 diabetes (T2D) and the related trait obesity. They found that the number of predicted candidate genes obtained by each of the different web tools varied dramatically and no match was found from the candidate lists between the two studies. Thornblad et al. (2007) tested three web tools (PosMed, GeneSniffer and SUSPECTS) on four disorders (breast cancer, Crohn's disease, age-related macular degeneration and schizophrenia) in which 10, 20 and 30 Mb segments of the chromosome containing the known susceptibility loci were tested. They found that the known disease gene(s) were not always in the top ranking

list and were more likely to rank higher when selected from a narrower genomic region. A combination of multiple candidate gene prioritization web tools was recommended and applied in most of the recent studies (Huang et al. 2008; Teber et al. 2009; Thornblad et al. 2007; Tiffin et al. 2006). Our data also showed that the priority lists obtained using our five web tools in concert were more enriched in genes with biological and nervous system function compared to prioritized genes using a single web tool Endeavour (93% vs. 89% for genes with biological functions,  $p = 0.6$ ; 31% vs. 19% for genes with nervous system functions,  $p = 0.21$ ; Table 5). However, as no statistically significant difference was found in the enrichments between the priority lists obtained using Endeavour and 5-tools in concert, the use of a single tool, such as Endeavour may still be considered, as it is more practical considering that creating a prioritization list using five web tools in concert is time consuming.

The effect of the training set on the priority list has not yet been investigated, nor was the difference between the

outcome of prioritization using random and disease-specific tests been addressed. It was surprising that the overlap between the prioritized gene lists obtained using random and ID training sets was very high and above  $>\sim 60\%$  (Table 4; Fig. 2). Two web tools (PosMed and Suspects), which use a disease term to create the “training” set internally, also resulted in a high overlap between prioritized genes obtained using random and ID training sets (39–60%) when terms “mental retardation”, “diabetes” or “breast cancer” were used to prioritize four larger CNVs (all contain  $>15$  genes/CNV) (data not shown). Despite the observed disease “non-specificity” of the web tools and training sets, our data show that candidate gene prioritization web tools in combination with OMIM training-gene set significantly enrich for genes with brain function, in comparison to non-prioritized genes, which could be useful for pre-selecting genes for further analysis from larger CNVs.

The aim of our tool comparison was not to identify the “best” prioritization tool, but to explore and apply them in light of the growing number of array-detected pathogenic CNVs, most of which have too many genes to allow for extensive literature searches individually. Currently, many of the available tools use different sources of information and algorithms for comparison with training genes and rely heavily on functional annotation of candidate genes. They are therefore inherently biased toward better characterized genes, which may result in less disease specificity (J. Gillis and P. Pavlidis, submitted). A user-friendly web tool which combines many datasets for identification of the most likely candidate ID genes is highly desirable in the current transition from conventional cytogenetics to ID-gene identification, as it can assist not only in exploring candidate disease genes for further molecular and functional analysis, but also in understanding the pathogenesis of ID.

Intellectual disability (ID) is a complex condition involving functional, developmental and structural alterations in the brain and/or nervous system. Despite hundreds of genes implicated in ID, the pathophysiological mechanisms of its origins remain largely unknown. In our study, we have compared the function of a series of prioritized candidate genes in de novo, familial and most common CNVs found in our ID cohort. A significant proportion ( $\sim 40\%$ , 26/66) of the prioritized genes in de novo deletion CNVs resulted in knockout mouse phenotypes manifesting abnormalities in nervous system development and are therefore likely involved in brain-related functions. This is consistent with a recently published report by Webber et al. (2009) in which a collection of 148 ID-associated de novo CNVs from the literature and  $>26,000$  benign CNVs from different sources (both encompassing  $>4,000$  genes) were assessed using MGI. ID-related CNVs (excluding benign CNVs) were found to be significantly enriched in genes

with nervous system phenotypes when disrupted in mice. The MGI web tool which catalogs the mouse knock out phenotype data is therefore a very informative web tool for disease gene study in humans.

For pathway and bioinformatics-based analyses of gene functions, we applied several different tools to minimize the bias from one single method. A different output was obtained from different pathway web tools, although most of them integrate KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway dataset (<http://www.genome.jp/kegg/pathway.html>). We noted that the prioritized genes from de novo CNVs are more involved in neuroactive ligand–receptor interaction and MAPK signaling pathway than expected by chance according to both WebGestalt and Pathway-Express. The involvement of MAPK pathways in ID-related genes have been observed recently (Sweatt 2001; Aoki et al. 2008) and 4/6 ID-related disease training-gene sets were also found to be enriched for genes from this pathway based on the Pathway-Express, Webgestalt and Gather web tool analysis. Among 10 genes from our de novo CNVs involved in these two pathways, five have already been reported in ID-related studies (*RPS6KAI* (Zeniou et al. 2002), *MEF2C* (Lipton et al. 2009), *CACNA1B* (Ladera et al. 2009), *MAPT* (Koolen et al. 2008) and *CRHR1* (Varela et al. 2006)). The remaining five genes are promising ID candidates with four demonstrating knockout mouse phenotypes related to nervous system abnormalities on MGI (*SIPR4* (Meng and Lee 2009), *OPRL1* (Manabe et al. 1998), *RASA1* (Henkemeyer et al. 1995) and *VIPR2* (Harmar et al. 2002)). The fifth gene, *UTS2R*, is a receptor of Urotensin-II which was found to act as a neurotransmitter in regulating various neurobiological activities including anxiety and depression in a recent study (do Rego et al. 2008). Other prioritized genes from our de novo CNVs that have been already found to be involved in ID include *CHRNA4* (Elghezal et al. 2007), *EHMT1* (Kleefstra et al. 2006), *FEZ1* (Lee et al. 2005), *KCNQ2* (Borgatti et al. 2004), *MBD5* (Jaillard et al. 2009), *MEN1* (Nakajima et al. 1999), *SOX8* (Pfeifer et al. 2000), *OTX1* (Laroche et al. 2008) and *SF1* (Schlaubitz et al. 2007). This suggests that, although further improvements for disease specificity are necessary, the candidate gene prioritization tools will remain a promising avenue for narrowing down functional genes harbored within pathogenic CNVs.

In conclusion, we applied array CGH and bioinformatics approaches to explore ID-related genes. Our results show that high-resolution array analysis with a combination of different computational approaches is helpful in extracting ID candidate genes and associating them with functional networks involved in ID. We believe that the opportunities to identify and prioritize the most likely candidate genes will facilitate their further molecular analysis and delineate their role in the pathogenesis of ID.

**Acknowledgments** This work was supported by funding from the Canadian Institutes for Health Research (CIHR) (MOP 74502; PI: ERS and MOP 64217; PI: MESL), Establishment Grant funding from the Michael Smith Foundation for Health Research (MESL) and an Ontario Mental Health Foundation grant (PI: JJA). PP is supported by a career award from the Michael Smith Foundation for Health Research, a CIHR New Investigator award, the Canadian Foundation for Innovation, and a Human Brain Project grant from the National Institutes of Health (GM076990). YQ is a trainee with the CIHR/NAAR-Autism Speaks STIHR Inter-Institute Autism Spectrum Disorders Training Program (PI: JJA). ERS is supported by a CIHR Institute of Genetics Clinician Investigator Award (2005–2009). MESL and ERS are Career Scholars supported by the Michael Smith Foundation for Health Research. The authors appreciate the collaboration and support of the participating subjects and their families.

## References

- Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 22:773–774
- Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol* 24:537–544
- Aoki Y, Niihori T, Narumi Y, Kure S, Matsubara Y (2008) The RAS/MAPK syndromes: novel roles of the RAS pathway in human genetic disorders. *Hum Mutat* 29:992–1006
- Borgatti R, Zucca C, Cavallini A, Ferrario M, Panzeri C, Castaldo P, Soldovieri MV, Baschiroto C, Bresolin N, Dalla Bernardina B, Tagliatalata M, Bassi MT (2004) A novel mutation in KCNQ2 associated with BFNC, drug resistant epilepsy, and mental retardation. *Neurology* 63:57–65
- Brymora A, Valova VA, Larsen MR, Roufogalis BD, Robinson PJ (2001) The brain exocyst complex interacts with RalA in a GTP-dependent manner: identification of a novel mammalian Sec3 gene and a second Sec15 gene. *J Biol Chem* 276:29792–29797
- Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucl Acids Res* 36:D724–D728
- Chang JT, Nevins JR (2006) GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics* 22:2926–2933
- Chelly J, Khelifaoui M, Francis F, Cherif B, Bienvenu T (2006) Genetics and pathophysiology of mental retardation. *Eur J Hum Genet* 14:701–713
- Chen J, Bardes EE, Aronow BJ, Jegga AG (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucl Acids Res* 37:W305–W311
- de Vries BB, White SM, Knight SJ, Regan R, Homfray T, Young ID, Super M, McKeown C, Splitt M, Quarrell OW, Trainer AH, Niermeijer MF, Malcolm S, Flint J, Hurst JA, Winter RM (2001) Clinical studies on submicroscopic subtelomeric rearrangements: a checklist. *J Med Genet* 38:145–150
- do Rego JC, Leprince J, Scalbert E, Vaudry H, Costentin J (2008) Behavioral actions of urotensin-II. *Peptides* 29:838–844
- Doh-ura K, Mekada E, Ogomori K, Iwaki T (2000) Enhanced CD9 expression in the mouse and human brains infected with transmissible spongiform encephalopathies. *J Neuropathol Exp Neurol* 59:774–785
- Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucl Acids Res* 31:3775–3781
- Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, Wijmenga C (2007) A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol Metab* 18:19–26
- Elghezal H, Hannachi H, Mougou S, Kammoun H, Triki C, Saad A (2007) Ring chromosome 20 syndrome without deletions of the subtelomeric and CHRNA4–KCNQ2 genes loci. *Eur J Med Genet* 50:441–445
- Eppig JT, Blake JA, Bult CJ, Richardson JE, Kadin JA, Ringwald M (2007) Mouse genome informatics (MGI) resources for pathology and toxicology. *Toxicol Pathol* 35:456–457
- Fan YS, Jayakar P, Zhu H, Barbooth D, Sacharow S, Morales A, Carver V, Benke P, Mundy P, Elsas LJ (2007) Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization. *Hum Mutat* 28:1124–1132
- Gibson WT, Harvard C, Qiao Y, Somerville MJ, Lewis ME, Rajcan-Separovic E (2008) Phenotype-genotype characterization of alpha-thalassemia mental retardation syndrome due to isolated monosomy of 16p13.3. *Am J Med Genet A* 146A:225–232
- Harmar AJ, Marston HM, Shen S, Spratt C, West KM, Sheward WJ, Morrison CF, Dorin JR, Piggins HD, Reubi JC, Kelly JS, Maywood ES, Hastings MH (2002) The VPAC(2) receptor is essential for circadian function in the mouse suprachiasmatic nuclei. *Cell* 109:497–508
- Harvard C, Malenfant P, Koochek M, Creighton S, Mickelson EC, Holden JJ, Lewis ME, Rajcan-Separovic E (2005) A variant Cri du Chat phenotype and autism spectrum disorder in a subject with de novo cryptic microdeletions involving 5p15.2 and 3p24.3–25 detected using whole genomic array CGH. *Clin Genet* 67:341–351
- Henkemeyer M, Rossi DJ, Holmyard DP, Puri MC, Mbamalu G, Harpal K, Shih TS, Jacks T, Pawson T (1995) Vascular system defects and neuronal apoptosis in mice lacking ras GTPase-activating protein. *Nature* 377:695–701
- Huang QY, Li GH, Cheung WM, Song YQ, Kung AW (2008) Prediction of osteoporosis candidate genes by computational disease-gene identification strategy. *J Hum Genet* 53:644–655
- Jaillard S, Dubourg C, Gerard-Blanluet M, Delahaye A, Pasquier L, Dupont C, Henry C, Tabet AC, Lucas J, Aboura A, David V, Benzacken B, Odent S, Pipiras E (2009) 2q23.1 microdeletion identified by array-CGH: an emerging phenotype with Angelman-like features? *J Med Genet* 46(12):847–855
- Kanehisa M, Bork P (2003) Bioinformatics in the post-sequence era. *Nat Genet* 33(Suppl):305–310
- Kleefstra T, Brunner HG, Amiel J, Oudakker AR, Nillesen WM, Magee A, Genevieve D, Cormier-Daire V, van Esch H, Fryns JP, Hamel BC, Sistermans EA, de Vries BB, van Bokhoven H (2006) Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. *Am J Hum Genet* 79:370–377
- Knight SJ, Regan R, Nicod A, Horsley SW, Kearney L, Homfray T, Winter RM, Bolton P, Flint J (1999) Subtle chromosomal rearrangements in children with unexplained mental retardation. *Lancet* 354:1676–1681
- Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 82:949–958
- Koolen DA, Sharp AJ, Hurst JA, Firth HV, Knight SJ, Goldenberg A, Saugier-veber P, Pfundt R, Vissers LE, Destree A, Grisart B, Rooms L, Van der Aa N, Field M, Hackett A, Bell K, Nowaczyk MJ, Mancini GM, Poddighe PJ, Schwartz CE, Rossi E, De Gregori M, Antonacci-Fulton LL, McLellan MD 2nd, Garrett JM, Wiechert MA, Miner TL, Crosby S, Ciccone R, Willatt L, Rauch A, Zenker M, Aradhya S, Manning MA, Strom TM, Wagenstaller J, Krepischi-Santos AC, Vianna-Morgante AM, Rosenberg C, Price SM, Stewart H, Shaw-Smith C, Brunner HG, Wilkie AO,

- Veltman JA, Zuffardi O, Eichler EE, de Vries BB (2008) Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. *J Med Genet* 45:710–720
- Koolen DA, Pfundt R, de Leeuw N, Hehir-Kwa JY, Nillesen WM, Neefs I, Scheltinga I, Sistermans E, Smeets D, Brunner HG, van Kessel AG, Veltman JA, de Vries BB (2009) Genomic microarrays in mental retardation: a practical workflow for diagnostic applications. *Hum Mutat* 30:283–292
- Kramer JM, van Bokhoven H (2009) Genetic and epigenetic defects in mental retardation. *Int J Biochem Cell Biol* 41:96–107
- Kumar RA, Sudi J, Babatz TD, Brune CW, Oswald D, Yen M, Nowak NJ, Cook EH, Christian SL, Dobyns WB (2009) A de novo 1p34.2 microdeletion identifies the synaptic vesicle gene RIMS3 as a novel candidate for autism. *J Med Genet* 47(2):81–90
- Ladera C, Martin R, Bartolome-Martin D, Torres M, Sanchez-Prieto J (2009) Partial compensation for N-type Ca(2+) channel loss by P/Q-type Ca(2+) channels underlines the differential release properties supported by these channels at cerebrocortical nerve terminals. *Eur J Neurosci* 29:1131–1140
- Laroche F, Ramoz N, Leroy S, Fortin C, Rousselot-Paillet B, Philippe A, Colleaux L, Bresson JL, Mogenet A, Golse B, Mouren-Simeoni MC, Gorwood P, Galli T, Simonneau M, Krebs MO, Robel L (2008) Polymorphisms of coding trinucleotide repeats of homeogenes in neurodevelopmental psychiatric disorders. *Psychiatr Genet* 18:295–301
- Lee S, Walker CL, Karten B, Kuny SL, Tennese AA, O'Neill MA, Wevrick R (2005) Essential role for the Prader-Willi syndrome protein necdin in axonal outgrowth. *Hum Mol Genet* 14:627–637
- Lee C, Iafrate AJ, Brothman AR (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* 39:S48–S54
- Letts VA, Valenzuela A, Dunbar C, Zheng QY, Johnson KR, Frankel WN (2000) A new spontaneous mouse mutation in the Kcne1 gene. *Mamm Genome* 11:831–835
- Lipton SA, Li H, Zaremba JD, McKercher SR, Cui J, Kang YJ, Nie Z, Soussou W, Talantova M, Okamoto S, Nakanishi N (2009) Autistic phenotype from MEF2C knockout cells. *Science* 323:208
- Liu XG, Liu YJ, Liu J, Pei Y, Xiong DH, Shen H, Deng HY, Papisian CJ, Drees BM, Hamilton JJ, Recker RR, Deng HW (2008) A bivariate whole genome linkage study identified genomic regions influencing both BMD and bone structure. *J Bone Miner Res* 23:1806–1814
- Macleod G, Dolle P, Petkovich M (2009) Genetic disruption of CYP26B1 severely affects development of neural crest derived head structures, but does not compromise hindbrain patterning. *Dev Dyn* 238:732–745
- Manabe T, Noda Y, Mamiya T, Katagiri H, Houtani T, Nishi M, Noda T, Takahashi T, Sugimoto T, Nabeshima T, Takeshima H (1998) Facilitation of long-term potentiation and memory in mice lacking nociceptin receptors. *Nature* 394:577–581
- Meng H, Lee VM (2009) Differential expression of sphingosine-1-phosphate receptors 1–5 in the developing nervous system. *Dev Dyn* 238:487–500
- Nakajima K, Sakurai A, Kubota T, Katai M, Mori J, Aizawa T, Fukushima Y, Hashizume K (1999) Multiple endocrine neoplasia type 1 concomitant with Prader-Willi syndrome: case report and genetic diagnosis. *Am J Med Sci* 317:346–349
- Osoegawa K, Vessere GM, Utami KH, Mansilla MA, Johnson MK, Riley BM, L'Heureux J, Pfundt R, Staaf J, van der Vliet WA, Lidral AC, Schoenmakers EF, Borg A, Schutte BC, Lammer EJ, Murray JC, de Jong PJ (2008) Identification of novel candidate genes associated with cleft lip and palate using array comparative genomic hybridisation. *J Med Genet* 45:81–86
- Pfeifer D, Poulat F, Holinski-Feder E, Kooy F, Scherer G (2000) The SOX8 gene is located within 700 kb of the tip of chromosome 16p and is deleted in a patient with ATR-16 syndrome. *Genomics* 63:108–116
- Qiao Y, Harvard C, Riendeau N, Fawcett C, Liu X, Holden JJ, Lewis ME, Rajcan-Separovic E (2008) Putatively benign copy number variants in subjects with idiopathic autism spectrum disorder and/or intellectual disability. *Cytogenet Genome Res* 123:79–87
- Qiao Y, Riendeau N, Koochek M, Liu X, Harvard C, Hildebrand MJ, Holden JJ, Rajcan-Separovic E, Lewis ME (2009) Phenomic determinants of genomic variation in autism spectrum disorders. *J Med Genet* 46(10):680–688
- Rajcan-Separovic E, Harvard C, Liu X, McGillivray B, Hall JG, Qiao Y, Hurlburt J, Hildebrand J, Mickelson EC, Holden JJ, Lewis ME (2007) Clinical and molecular cytogenetic characterisation of a newly recognised microdeletion syndrome involving 2p15-16.1. *J Med Genet* 44:269–276
- Rajcan-Separovic E, Qiao Y, Tyson C, Harvard C, Fawcett C, Kalousek D, Stephenson M, Philipp T (2010) Genomic changes detected by array CGH in human embryos with developmental defects. *Mol Hum Reprod* 16:125–134
- Roeleveld N, Zielhuis GA, Gabreels F (1997) The prevalence of mental retardation: a critical review of recent literature. *Dev Med Child Neurol* 39:125–132
- Schlaubitz S, Yatsenko SA, Smith LD, Keller KL, Vissers LE, Scott DA, Cai WW, Reardon W, Abdul-Rahman OA, Lammer EJ, Lifchez CA, Magenis E, Veltman JA, Stankiewicz P, Zabel BU, Lee B (2007) Ovotestes and XY sex reversal in a female with an interstitial 9q33.3–q34.1 deletion encompassing NR5A1 and LMX1B causing features of Genitopatellar syndrome. *Am J Med Genet A* 143A:1071–1081
- Stein LD (2003) Integrating biological databases. *Nat Rev Genet* 4:337–345
- Sweatt JD (2001) The neuronal MAP kinase cascade: a biochemical signal integration system subserving synaptic plasticity and memory. *J Neurochem* 76:1–10
- Tebar LA, Geranton SM, Parsons-Perez C, Fisher AS, Bayne R, Smith AJ, Turmaine M, Perez-Luz S, Sheasby A, De Felipe C, Ruff C, Raivich G, Hunt SP (2008) Deletion of the mouse RegIIIbeta (Reg2) gene disrupts ciliary neurotrophic factor signaling and delays myelination of mouse cranial motor neurons. *Proc Natl Acad Sci USA* 105:11400–11405
- Teber ET, Liu JY, Ballouz S, Fatkin D, Wouters MA (2009) Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies. *BMC Bioinformatics* 10(Suppl 1):S69
- Thornblad TA, Elliott KS, Jowett J, Visscher PM (2007) Prioritization of positional candidate genes using multiple web-based software tools. *Twin Res Hum Genet* 10:861–870
- Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, Adeyemo A, Patti ME, Semple CA, Hide W (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucl Acids Res* 34:3067–3081
- Tiffin N, Okpechi I, Perez-Iratxeta C, Andrade-Navarro MA, Ramasar R (2008) Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes. *Physiol Genomics* 35:55–64
- Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucl Acids Res* 36:W377–W384
- Tyson C, McGillivray B, Chijiwa C, Rajcan-Separovic E (2004) Elucidation of a cryptic interstitial 7q31.3 deletion in a patient with a language disorder and mild mental retardation by array-CGH. *Am J Med Genet A* 129A:254–260

- Tyson C, Harvard C, Locker R, Friedman JM, Langlois S, Lewis ME, Van Allen M, Somerville M, Arbour L, Clarke L, McGilivray B, Yong SL, Siegel-Bartel J, Rajcan-Separovic E (2005) Submicroscopic deletions and duplications in individuals with intellectual disability detected by array-CGH. *Am J Med Genet A* 139:173–185
- van Vlijmen T, Vleugel M, Evers M, Mohammed S, Wulf PS, Heck AJ, Hoogenraad CC, van der Sluijs P (2008) A unique residue in rab3c determines the interaction with novel binding protein Zwint-1. *FEBS Lett* 582:2838–2842
- Varela MC, Krepischi-Santos AC, Paz JA, Knijnenburg J, Szuhai K, Rosenberg C, Koiffmann CP (2006) A 17q21.31 microdeletion encompassing the MAPT gene in a mentally impaired patient. *Cytogenet Genome Res* 114:89–92
- Vissers LE, van Ravenswaaij CM, Admiraal R, Hurst JA, de Vries BB, Janssen IM, van der Vliet WA, Huys EH, de Jong PJ, Hamel BC, Schoenmakers EF, Brunner HG, Veltman JA, van Kessel AG (2004) Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat Genet* 36:955–957
- Vissers LE, Veltman JA, van Kessel AG, Brunner HG (2005) Identification of disease genes by whole genome CGH arrays. *Hum Mol Genet* 14(Spec no. 2):R215–R223
- von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucl Acids Res* 35:D358–D362
- Webber C, Hehir-Kwa JY, Nguyen DQ, de Vries BB, Veltman JA, Ponting CP (2009) Forging links between human mental retardation-associated CNVs and mouse gene knockout models. *PLoS Genet* 5:e1000531
- Yonan AL, Palmer AA, Smith KC, Feldman I, Lee HK, Yonan JM, Fischer SG, Pavlidis P, Gilliam TC (2003) Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction. *Genes Brain Behav* 2:303–320
- Yoshida Y, Makita Y, Heida N, Asano S, Matsushima A, Ishii M, Mochizuki Y, Masuya H, Wakana S, Kobayashi N, Toyoda T (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res* 37:W147–W152
- Zeniou M, Ding T, Trivier E, Hanauer A (2002) Expression analysis of RSK gene family members: the RSK2 gene, mutated in Coffin-Lowry syndrome, is prominently expressed in brain structures essential for cognitive function and learning. *Hum Mol Genet* 11:2929–2940
- Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucl Acids Res* 33:W741–W748