Identification and analysis of novel genes expressed in the mouse embryonic facial primordia

Jennifer S. Bennetts 1, Lindsay F. Fowles 1, Natalie C. Butterfield 1, Jennifer L. Berkman 1, Rohan D. Teasdale 1,2, Fiona Simpson 1, and Carol Wicking 1

1 Institute for Molecular Bioscience and 2ARC Centre in Bioinformatics, The University of Queensland, QLD 4072, Australia

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and Methods
   3.1. Bioinformatics
   3.2. Whole mount in situ hybridisation
   3.3. Cloning of epitope-tagged coding sequences for candidate genes
   3.4. Cell culture and immunofluorescence
4. Results
   4.1. Initial candidates identified by sequence analysis from the developmentally enriched library
   4.2. Expression analysis by whole mount in situ hybridisation
   4.3. In-depth in silico analysis
   4.4. Cloning into expression vectors and transfection into mammalian cells
5. Discussion
   5.1. Whole mount in situ expression analysis and bioinformatics to identify novel/poorly characterised genes with development roles
   5.2. Prediction of domains, subcellular localisation and transfection of tagged constructs
   5.3. Potential association of novel and poorly characterised molecules with human hereditary disorders
6. Conclusions and Summary
7. Acknowledgments
8. References

1. ABSTRACT

Craniofacial anomalies are a common feature of human congenital dysmorphology syndromes, suggesting that genes expressed in the developing face are likely to play a wider role in embryonic development. To facilitate the identification of genes involved in embryogenesis, we previously constructed an enriched cDNA library by subtracting adult mouse liver cDNA from that of embryonic day (E)10.5 mouse pharyngeal arch cDNA. From this library, 273 unique clones were sequenced and known proteins binned into functional categories in order to assess enrichment of the library (1). We have now selected 31 novel and poorly characterised genes from this library and present bioinformatic analysis to predict proteins encoded by these genes, and to detect evolutionary conservation. Of these genes 61% (19/31) showed restricted expression in the developing embryo, and a subset of these was chosen for further in silico characterisation as well as experimental determination of subcellular localisation based on transient transfection of predicted full-length coding sequences into mammalian cell lines. Where a human orthologue of these genes was detected, chromosomal localisation was determined relative to known loci for human congenital disease.

2. INTRODUCTION

Embryonic development requires the highly coordinated expression of a host of genes in a precise spatio-temporal manner. Mutations in specific genes contributing to this program of expression can result in a range of dysmorphology syndromes in humans. Craniofacial abnormalities are a common feature of such syndromes, and those disorders that present with craniofacial defects together with abnormalities in other organs and tissues suggest that genes involved in facial development may also play a significant role in the development of other structures. For example, both craniofacial and limb malformations in humans or mice have been associated concurrently with mutations in genes such as TWIST (2), fibroblast growth factor receptor 2 (FGFR2) (3) and GLI3 (4). To date, genes responsible for only a fraction of human congenital abnormalities have been identified, and the discovery of novel genes involved in normal embryonic development will provide further candidates for such disease genes, as well as contributing to our understanding of basic developmental processes.

In order to identify genes with a role in embryonic development, we performed a screen to uncover
genes expressed in mouse embryonic facial primordia, and have previously reported those known genes identified in this screen (1). Briefly, a subtractive hybridisation approach was employed in order to isolate genes expressed specifically in the E10.5 first and second pharyngeal arches. Subtraction was performed against adult mouse liver cDNA to reduce representation of ubiquitously expressed genes. Analysis of the subtracted library revealed successful enrichment for developmentally expressed genes with important functional roles in embryogenesis, and hence validated our approach. We have now selected either completely novel or poorly characterised candidates for further analysis.

In total, 453 clones from the subtracted cDNA library were randomly picked and sequenced. The insert sequence from each of these clones was analysed by BLAST search for homologies against sequences within the publicly available databases, and 273 of these clones were shown to represent unique predicted transcripts, the majority of which corresponded to known genes (1). Of these 273, nine sequences were identified which were identical to novel/predicted gene transcripts, four were identical to transcripts which were annotated based on their predicted protein sequence homology to known proteins, six corresponded to identified/published molecules without characterised functional roles, ten proteins, six corresponded to identified/published on their predicted protein sequence homology to known genes (1). Contiguous EST (expressed sequence tag) sequence assemblies were generated using Sequencer 4.2 (Gene Codes, MI, USA) and analysed as previously described to obtain predicted full-length cDNA sequences (1). In addition, the recently released FANTOM III (7) sequence database (http://fantom3.gsc.riken.jp/db) was used to determine and cross-reference sequences. Analysis of predicted protein sequences for the presence of functional protein domains was performed as described previously (1). Additionally, the databases Interpro (http://www.ebi.ac.uk/InterProScan/), Prosite (http://au.expasy.org/prosite/), PsortII (http://psort.nibb.ac.jp/form2.html) and the LOCATE Subcellular Localisation Database (http://locate.imb.uq.edu.au/) were used to determine putative domains and motifs, and only predictions which were above the default threshold value for each program are reported. Human chromosomal localisations were obtained by BLAST search against the publicly available human genome database at (http://www.ncbi.nlm.nih.gov/omim/ ). Human disease loci mapping to each region were obtained using the GenAtlas (http://www.dsi.univ-paris5.fr/genatlas/) and OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) databases. To analyse relative representation of expressed transcripts in the adult mouse, the Source database (http://smd.stanford.edu/cgi-bin/source/sourceSearch) was utilised (8).

3. Whole mount in situ hybridisation

Whole mount in situ hybridisation was performed as described previously (1, 9). Briefly, E10.5 embryos were dissected from pregnant CD1 female mice following timed matings, fixed overnight in 4% parafomaldehyde/PBS at 4°C, dehydrated and then rehydrated through a methanol series. Embryos were placed in prehybridisation buffer and incubated 2 hours to overnight at 65°C. Digoxigenin (DIG)-labelled riboprobes were transcribed from linearised clones using T7 or SP6 polymerase and added to prehybridised embryos at a concentration of 0.2–1.0 µg/ml. Both antisense and sense probes were examined. After overnight hybridisation, embryos were washed, blocked and incubated with anti-DIG antibody conjugated to alkaline phosphatase (Roche Diagnostics, Mannheim, Germany). After a further round of washing, embryos were incubated with the colour reagents nitro blue tetrazolium (Roche Diagnostics) and 5-bromo-4-chloro-3-indolylphosphate (Roche Diagnostics) until satisfactory colour development was achieved. Background was reduced by washing in PBS containing 1% Triton X-100 at 4°C for one to three days and the colour then fixed with 4% parafomaldehyde in PBS overnight at 4°C.

3. MATERIALS AND METHODS

3.1. Bioinformatics

Homology searches were performed using BLAST algorithms (5, 6) to identify homologous nucleotide sequences through the National Center for Biotechnology Information (NCBI) public database (http://www.ncbi.nlm.nih.gov/BLAST/). Additionally the NCBI HomoloGene tool was used to identify orthologues of protein sequences in a range of species (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene&cmd=search&term=). Contiguous EST (expressed sequence tag) sequence assemblies were generated using Sequencer 4.2 (Gene Codes, MI, USA) and analysed as previously described to obtain predicted full-length cDNA sequences (1). In addition, the recently released FANTOM III (7) sequence database (http://fantom3.gsc.riken.jp/db) was used to determine and cross-reference sequences. Analysis of predicted protein sequences for the presence of functional protein domains was performed as described previously (1). Additionally, the databases Interpro (http://www.ebi.ac.uk/InterProScan/), Prosite (http://au.expasy.org/prosite/), PsortII (http://psort.nibb.ac.jp/form2.html) and the LOCATE Subcellular Localisation Database (http://locate.imb.uq.edu.au/) were used to determine putative domains and motifs, and only predictions which were above the default threshold value for each program are reported. Human chromosomal localisations were obtained by BLAST search against the publicly available human genome database at (http://www.ncbi.nlm.nih.gov/omim/ ). Human disease loci mapping to each region were obtained using the GenAtlas (http://www.dsi.univ-paris5.fr/genatlas/) and OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) databases. To analyse relative representation of expressed transcripts in the adult mouse, the Source database (http://smd.stanford.edu/cgi-bin/source/sourceSearch) was utilised (8).
Novel genes in the mouse embryonic facial primordia

3.3. Cloning of epitope-tagged coding sequences for candidate genes

Where available, RIKEN clones containing full-length predicted coding sequences (supplied from the RIKEN mouse cDNA clone set FANTOM I and II, established by the Genome Exploration Group, RIKEN GSC [http://genome.riken.go.jp/] and replicated and provided by K. K. DNAform (10-13)) were obtained from the SRC microarray facility at the Institute for Molecular Bioscience. The use of these clones was sublicensed from The Institute of Physical and Chemical Research (RIKEN). If suitable clones were not available, mRNA was extracted from whole E10.5 mouse embryos using the RNeasy Kit (Qiagen, CA, USA), and cDNA was generated using MMLV-Reverse Transcriptase (Invitrogen Life Technologies, Carlsbad, CA, USA) according to the manufacturer’s instructions. RIKEN clone or cDNA templates were used to amplify predicted full-length protein-coding regions of each candidate gene by polymerase chain reaction (PCR) using flanking primers (Genernkeys) with engineered restriction sites. The PCR products were each cloned into pGEMTeasy (Promega, Madison, WI, USA) using standard techniques (14), sequenced (ABI PRISM BigDye™ Terminator Version 3.1) and analysed (Australian Genome Research Facility). These constructs were digested using appropriate restriction enzymes (New England Biolabs Inc., Beverley, MA, USA) to release the full-length coding sequences which were then ligated into a pcDNA3 (Invitrogen) vector 5' to a HA (hemagglutinin) or myc epitope tag allowing over-expression of carboxy-terminally tagged predicted full-length proteins in cultured mammalian cells.

3.4. Cell culture and immunofluorescence

Epitope-tagged constructs were transiently transfected into BHK-21 (Syrian golden hamster kidney) and HeLa (human cervical adenocarcinoma) cells using Lipofectamine 2000 (Invitrogen). Prior to transfection, cells were cultured on glass coverslips in multi-well dishes containing DMEM (Invitrogen) supplemented with 5-10% fetal calf serum. Where necessary for visualisation, cells were incubated with 0.05% saponin in PBS for 60 seconds to wash out the cytosol. Cells were fixed 24-48 hours after transfection in 3% paraformaldehyde in PBS for 15-30 minutes. Fixed cells were permeabilised with 0.1% Triton X-100 in PBS for 5-10 minutes, washed and blocked with 2% bovine serum albumin in PBS for a minimum of 30 minutes, and then incubated with appropriate primary antibodies diluted according to the manufacturer’s instructions in blocking solution for 60 minutes. Primary antibodies used were: monoclonal anti-HA 16B12 (Santa Cruz Biotechnology Inc., CA, USA), polyclonal anti-HA (Cell Signaling Technology, Beverley MA, USA), monoclonal anti-myc 9B11 (Cell Signaling), polyclonal anti-myc (Cell Signaling). For colocalisation experiments cells were incubated with MitoTracker Red (Molecular Probes, Eugene, OR, USA), monoclonal anti-human vinculin antibody V9131 (Sigma, St Louis, MO, USA), or co-transfected with the construct I-CAT-GFP (15). Cells were washed with blocking solution for 30 minutes, and incubated with anti-mouse or anti-rabbit Alexa fluor secondary antibodies (Molecular Probes) diluted in blocking solution for 60 minutes. Cells were washed for 30 minutes in blocking solution and incubated with DAPI nuclear stain, then washed with PBS and mounted with Mowiol 4-88 (Calbiochem, La Jolla, CA, USA). All incubations and washes were performed at room temperature. Cells were visualised using an Olympus AX-70 microscope (Tokyo, Japan) with a 100X UplanApo oil immersion objective, and images were captured with a DAGE-MTI CCD300-RC ccd camera (Michigan, IN, USA) using NIH Image 1.62. Merged images were produced using Adobe Photoshop 7.0. For colocalisation figures, cells were visualised using an Axiovert 200 M SP LSM 510 META confocal laser scanning microscope (Zeiss, Germany) under oil with a 63X or 100X objective, at excitation and emission wavelengths of 488 and 543, 505-530 and ≥560 nm for green and red fluorescence, respectively. Data was processed using LSM 510 META (Zeiss) software and images were assembled using Adobe Photoshop 7.0.

4. RESULTS

4.1. Initial candidates identified by sequence analysis from the developmentally enriched library

From a total of 273 uniquely represented sequences from our subtracted cDNA library (1), 31 non-redundant sequences were selected for further study. These were identical to regions of nucleotide sequences corresponding to novel or poorly characterised molecules based on homology comparisons by BLAST search against the publicly available databases. Overlapping contiguous EST sequence assemblies were generated to obtain and confirm predicted full-length transcripts. Open reading frames were identified, and these molecules were categorised into one of five groups (Table 1). Nine represented hypothetical/predicted novel protein coding genes (nv1-nv9); three were novel genes with annotations based on protein sequence homology to either putative orthologues in other species (nv10-nv11) or to characterised proteins with similar domain structure (Trp53inp2); six represented molecules which had previously been published either as part of a large-scale screen or based on expression data, with no characterised functional role (Etea, Kctd3, Cdc4, Cxxl, Hnf1 and Fio); ten represented molecules which were poorly characterised and had not been previously described as developmentally associated (Usp52, p15^Paf, Elp3, Cdc7, Limd1, C-mir, Thoc3, Maged2, Ras11b and Psf2); and two represented transcripts for which a confidently predicted open reading frame could not be identified (nv12-nv13), either because of a lack of available sequence data in the database, or because they may potentially represent either pseudogenes or regulatory antisense RNA transcripts. At the time that this work was carried out, for one sequence an open reading frame could not be identified, and this gene was not analysed further. However recently released FANTOM III sequence data (7) has allowed us to determine that this sequence actually corresponds to the 3' UTR of the predicted protein-coding gene Zcchc3 (NM_175126).

4.2. Expression analysis by whole mount in situ hybridisation

To assess the likelihood of each of these genes playing a role in embryogenesis, the pattern of expression
Novel genes in the mouse embryonic facial primordia

<table>
<thead>
<tr>
<th>Name</th>
<th>Gene accession</th>
<th>Entrez gene/locus</th>
<th>Protein name/description</th>
<th>Human products</th>
<th>General comments</th>
<th>Status</th>
<th>isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td>p15</td>
<td>NM_001022662</td>
<td>6330403K20Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>Novel</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m2</td>
<td>AK355457</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m3</td>
<td>AK355457</td>
<td>1520194C08Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC66172</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m4</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m5</td>
<td>AK041293</td>
<td>1009003L08Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC66172</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m6</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m7</td>
<td>AK041293</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC66172</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m8</td>
<td>AK355457</td>
<td>1520194C08Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC66172</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m9</td>
<td>AK355457</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC66172</td>
<td>N</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 1. Novel poorly characterised molecules selected from subtracted cDNA library for further analysis**

<table>
<thead>
<tr>
<th>Name</th>
<th>Entrez gene/locus</th>
<th>4531195H14Rik</th>
<th>Protein name/description</th>
<th>Human products</th>
<th>General comments</th>
<th>Status</th>
<th>isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td>m10</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m11</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m12</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m13</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 2. Mouse genes with known roles in development**

<table>
<thead>
<tr>
<th>Name</th>
<th>Entrez gene/locus</th>
<th>4531195H14Rik</th>
<th>Protein name/description</th>
<th>Human products</th>
<th>General comments</th>
<th>Status</th>
<th>isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td>m14</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m15</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m16</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m17</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 3. Molecules poorly characterised and with known roles in development**

<table>
<thead>
<tr>
<th>Name</th>
<th>Entrez gene/locus</th>
<th>4531195H14Rik</th>
<th>Protein name/description</th>
<th>Human products</th>
<th>General comments</th>
<th>Status</th>
<th>isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td>m18</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m19</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 4. Molecules with characterised functional role**

<table>
<thead>
<tr>
<th>Name</th>
<th>Entrez gene/locus</th>
<th>4531195H14Rik</th>
<th>Protein name/description</th>
<th>Human products</th>
<th>General comments</th>
<th>Status</th>
<th>isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td>m20</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m21</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m22</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m23</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 5. Molecules with partially conserved orthologous role in development**

<table>
<thead>
<tr>
<th>Name</th>
<th>Entrez gene/locus</th>
<th>4531195H14Rik</th>
<th>Protein name/description</th>
<th>Human products</th>
<th>General comments</th>
<th>Status</th>
<th>isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td>m24</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m25</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m26</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
<tr>
<td>m27</td>
<td>NM_002455</td>
<td>2610507P14Rik</td>
<td>SPOC2</td>
<td>Hypothetical protein</td>
<td>LOC65649</td>
<td>N</td>
<td>✓</td>
</tr>
</tbody>
</table>

A “✓” represents molecules for which spatially-restricted expression was observed by whole mount in situ hybridisation in the developing mouse embryo at E10.5. Abbreviations: MGI (Mouse Genome Informatics); N, novel; ND, not detected or not able to be obtained through public databases; nv, novel molecule; ORF, open reading frame; U, uncharacterised role in embryonic development.

was determined by whole mount in situ hybridisation in E10.5 mouse embryos. DIG-labelled RNA probes were transcribed from the subtracted cDNA library clone templates, and sense probes were also hybridised as negative controls. Of these 31 candidates, 19 exhibited spatially restricted expression in the developing embryo, and 17 of these are shown in Figure 1. Of the remaining two, further characterisation of one of these genes (p15\(^{\text{ov}}\)) has recently been reported (16), and the other (Cdc45) is under further investigation and will be reported at a later date. In addition to expression in the facial primordia, each of these 19 genes was also shown to be expressed in the developing limbs, providing support for the concept that conserved molecular processes govern both limb and craniofacial development (17). These genes were also expressed in a number of other regions most commonly including the heart, otic vesicle, developing neural tissues, developing tail, somites, and mesonephros (Figure 1). Expression of each of these genes in adult mouse tissues was analysed in silico using the SOURCE database.
Figure 1. A selection of novel/poorly characterised genes from the subtracted cDNA library which showed restricted expression in E10.5 mouse embryos. Expression patterns for the following genes are shown: *nv1* (A), *nv2* (B), *nv4* (C), *nv5* (D), *nv6* (E), *nv9* (F), *nv11* (G), *Trp53inp2* (H), *Zcchc3* (I), *Hn1* (J), *Fto* (K), *Elp3* (L), *Limd1* (M), *Thoc3* (N), *Maged2* (O), *Psf2* (P) and *nv13* (Q). Common sites of expression to all genes examined include the maxillary and mandibular processes, the second pharyngeal arch, the frontonasal process and the developing fore- and hindlimb buds (A-Q). Other sites of expression include the otic vesicle (developing ear) (A, B, D-G, I, J, L-N, Q), the developing fore-, mid- and hindbrain (A-J, L-P), the somites (B, C, K, L, P, Q), the developing heart (A, D, E, H, J, M), the dorsal root ganglia (D, H) and the spinal chain ganglia (H), the developing tail (A-C, K, O, I), the tailbud (Q) and the mesonephros (G, I, M, P). Abbreviations: drg, dorsal root ganglion; fb, forebrain; fl, forelimb; fnp, frontonasal process; h, heart; hb, hindbrain; hl, hindlimb; mb, midbrain; md, mandibular process; mn, mesonephros; mx, maxillary process; ov, otic vesicle; pa2, 2nd pharyngeal arch; s, somite; scg, spinal chain ganglion; t, developing tail; tb, tailbud.
Novel genes in the mouse embryonic facial primordia

(http://smd.stanford.edu/cgi-bin/source/sourceSearch) and the majority were found to have widespread expression in the adult (data not shown). This is consistent with a number of known developmental genes, such as those encoding certain members of the transforming growth factor and Wnt signalling molecule families (18, 19), which play very specific and restricted roles in the embryo, but are expressed more widely in the adult and may carry out different functions.

4.3. In depth in silico analysis

For each of the expression patterns shown in Figure 1, the corresponding gene was analysed more extensively using in silico bioinformatics techniques, except for \(nv13\), for which a confidently predicted open reading frame was not identified, and \(nv9\) (\(AK141733\)), for which there was insufficient sequence data in the public databases to extend the 5’ end of the transcript, and obtain a full-length sequence. Protein sequences were analysed using publicly available domain and motif prediction tools to identify putative functional regions and subcellular distribution (Table 2).

Molecules with important roles in embryonic development are likely to be conserved across species during evolution (20). For each molecule, BLAST and HomoloGene searches were used to identify potential orthologous amino acid sequences in a range of commonly annotated species within the NCBI sequence database (Table 2). Additionally, for each gene for which a human orthologue was identified the human chromosomal localisation was determined by BLAST search against the human genome database at NCBI. To investigate potential association of genes with human hereditary disorders and developmental phenotypes, diseases which show linkage to each chromosomal region were established using the GenAtlas and OMIM databases (Table 2).

4.4. Cloning into expression vectors and transfection into mammalian cells

In order to assess the subcellular localisation of selected proteins, the mouse predicted open reading frame for each molecule was cloned into a C-terminally HA or myc epitope-tagged expression vector. Constructs were transiently transfected into the mammalian cell lines HeLa and BHK-21 and analysed by immunofluorescence with an anti-HA or anti-myc antibody. Immunofluorescence analysis of fixed transfected cells revealed the subcellular localisation of each transfected protein, and showed a high correlation of localisation in both cell lines analysed (Figure 2, Figure 3). \(Nv4\) (Figure 2E,F), \(Nv5\) (Figure 2G,H) and Fatso (Figure 3C,D) showed a predominantly nuclear localisation, and \(Nv2\) (Figure 2C,D), \(Nv6\) (Figure 2I,J), \(Hn1\) (Figure 3A,B) and \(Psf2\) (Figure 3M,N) showed a nuclear and diffuse cytoplasmic distribution. \(Nv1\) (Figure 2A,B) was localised to what appeared to be endoplasmic reticulum, while \(Nv11\) (Figure 2K,L) and \(Elp3\) (Figure 3E,F) were localised to the nucleus as well as what appeared to be mitochondria. \(Trp53imp2\) (Figure 2M,N) was distributed within the nucleus and in puncta surrounding the nucleus, and \(Limd1\) (Figure 3G,H) localised throughout the cytoplasm and also at focal adhesions at the cell surface. Overexpression of Thoc3 (Figure 3I,J) caused aggregation in structures presumed to be aggresomal as well as localisation to vesicular-like structures within the cell, and \(Maged2\) (Figure 3K,L) localised to the nucleus (specifically within nucleoli), endosomal-like structures and at the plasma membrane. To confirm preliminary data, the localisation of a subset of these molecules was explored further by colocalisation with known subcellular markers using confocal microscopy (Figure 4). \(Nv1\) was co-transfected with \(I-CAT\)-\(GFP\) and colocalisation confirmed an endoplasmic reticulum distribution (Figure 4A-C). \(Nv11\) and \(Elp3\) staining colocalised with MitoTracker Red in mitochondria (Figure 4D-I), and \(Limd1\) localised with vinculin at focal adhesions (Figure 4J-L). Subcellular localisation data is summarised in Table 2 alongside PsortII predicted localisations.

5. DISCUSSION

5.1. Whole mount in situ expression analysis and bioinformatics to identify novel/poorly characterised genes with developmental roles

We previously reported the generation and analysis of a subtracted cDNA library which was shown to be significantly enriched for molecules with functions in embryogenesis (1). Of a total of 273 non-redundant sequences analysed from this library, 31 of these (11.4%) were identified as novel/poorly characterised candidates and prioritised for further analysis to investigate potential developmental roles.

Of these 31 genes, 19 were shown to be expressed in a restricted pattern in the developing mouse embryo and of these, 17 were selected for further characterisation in the current study. For these genes, a range of expression patterns was observed. Notably, each exhibited staining in both the craniofacial region and the limbs, which lends further support for the concept that genes involved in craniofacial development are also likely to be implicated in limb development. Furthermore, within these structures a variety of expression patterns were observed. For example, while some genes appeared to be expressed generally throughout the mesenchymal core of the pharyngeal arches, others showed more restricted expression to distal mesenchyme and ectoderm (eg. \(Ffo\), Figure 1K). This may reflect the fact that patterning and outgrowth of the facial primordia require epithelial-mesenchymal signalling interactions (reviewed in (21)), and those more restricted genes may potentially be involved in these processes. Also, some genes appeared to show restricted expression within the first pharyngeal arch, and were expressed at lower levels in the second arch (eg. \(Hn1\), Figure 1J). These differences in expression may signify molecular differences between the arches, and may also suggest potential pathways in which these uncharacterised molecules could function. In the limb a range of restricted expression domains were also observed, consistent with known reciprocal molecular interactions governing patterning and development in all three axes of the limb (22).
# Novel genes in the mouse embryonic facial primordia

## Table 2. Summary of further analysis of predicted proteins encoded by developmentally restricted novel/ poorly characterised genes

<table>
<thead>
<tr>
<th>Name (aa)</th>
<th>Functional Domain/ Motifs (amino acid sequence/ position)</th>
<th>Predicted domain (amino acid sequence/ position)</th>
<th>Subcellular distribution of construct</th>
<th>Orthology identification</th>
<th>Ac. ID, species/ % aa identity</th>
<th>Human locus</th>
<th>Disease/abnormality and genomic region implicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>nv1 (121aa)</td>
<td>Transmembrane (96-118)</td>
<td>69.8% nuclear</td>
<td>Endoplasmic reticulum</td>
<td>ND</td>
<td>17p13</td>
<td>ND</td>
<td></td>
</tr>
<tr>
<td>nv2 (195aa)</td>
<td>ND</td>
<td>47.8% cytoplasmic, NLS (41)</td>
<td>Nucleus, cytoplasm</td>
<td>CAI01428.g (97%)</td>
<td>NP_959013 (85%)</td>
<td>CAI05669.m (84%)</td>
<td>17q21</td>
</tr>
<tr>
<td>nv3 (112aa)</td>
<td>ND</td>
<td>73.9% nuclear</td>
<td>Nucleus</td>
<td>XP_573678.m (96%)</td>
<td>NP_01102408.m (67%)</td>
<td>CAI07058.m (63%)</td>
<td>19p13</td>
</tr>
<tr>
<td>nv4 (64aa)</td>
<td>ND</td>
<td>7 residue NLS (248)</td>
<td>Nucleus</td>
<td>NP_00107948.r (98%)</td>
<td>NP_956900 (60%)</td>
<td>AAH79002.m (67%)</td>
<td>CAI02953.m (57%)</td>
</tr>
<tr>
<td>nv5 (211aa)</td>
<td>ND</td>
<td>34.8% cytoplasmic, NLS (248)</td>
<td>Nucleus, mitochondria</td>
<td>NP_00014210.o (98%)</td>
<td>XP_417831.g (98%)</td>
<td>AAH72849.m (83%)</td>
<td>CAI03584.m (71%)</td>
</tr>
<tr>
<td>nv6 (165aa)</td>
<td>ND</td>
<td>30.4% mitochondrial, 17.4% nuclear</td>
<td>Nucleus, mitochondria</td>
<td>XP_252540.g (98%)</td>
<td>XP_825240.g (98%)</td>
<td>NA</td>
<td>20q11</td>
</tr>
<tr>
<td>nv7 (65aa)</td>
<td>ND</td>
<td>62.5% cytoplasmic</td>
<td>Nucleus, perinuclear region</td>
<td>CAI05290.m (98%)</td>
<td>XP_358010.r (97%)</td>
<td>NP_959176.m (67%)</td>
<td>NP_900725.m (74%)</td>
</tr>
<tr>
<td>nv8 (371aa)</td>
<td>ND</td>
<td>30.4% cytoplasmic</td>
<td>Nucleus, mitochondria</td>
<td>XP_358010.r (97%)</td>
<td>NP_959176.m (67%)</td>
<td>NP_900725.m (74%)</td>
<td>20p12.1-16p12</td>
</tr>
<tr>
<td>nv9 (52aa)</td>
<td>ND</td>
<td>34.8% nuclear</td>
<td>Nucleus</td>
<td>AAH79002.m (67%)</td>
<td>CAI05424.m (40%)</td>
<td>NP_00105803.m (93%)</td>
<td>XP_533120</td>
</tr>
<tr>
<td>nv10 (101aa)</td>
<td>ND</td>
<td>34.8% nuclear</td>
<td>Nucleus</td>
<td>AAH79002.m (67%)</td>
<td>XP_533120</td>
<td>NP_00105803.m (93%)</td>
<td>XP_533120</td>
</tr>
<tr>
<td>nv11 (65aa)</td>
<td>ND</td>
<td>34.8% cytoplasmic</td>
<td>Nucleus, mitochondria</td>
<td>NP_00014210.o (98%)</td>
<td>XP_417831.g (98%)</td>
<td>AAH72849.m (83%)</td>
<td>CAI03584.m (71%)</td>
</tr>
<tr>
<td>nv12 (155aa)</td>
<td>ND</td>
<td>30.4% mitochondrial, 17.4% nuclear</td>
<td>Nucleus, mitochondria</td>
<td>XP_358010.r (97%)</td>
<td>NP_959176.m (67%)</td>
<td>NP_900725.m (74%)</td>
<td>20p12.1-16p12</td>
</tr>
<tr>
<td>nv13 (35aa)</td>
<td>ND</td>
<td>60.9% nuclear, 21.7% cytoplasmic</td>
<td>Cytoplasm, focal adhesions</td>
<td>XP_250734.r (92%)</td>
<td>XP_541912.r (78%)</td>
<td>XP_418803.m (83%)</td>
<td>XP_691259.d (60%)</td>
</tr>
<tr>
<td>nv14 (366aa)</td>
<td>ND</td>
<td>30.4% cytoplasmic</td>
<td>Nucleus, mitochondria</td>
<td>XP_358010.r (97%)</td>
<td>NP_959176.m (67%)</td>
<td>NP_900725.m (74%)</td>
<td>20p12.1-16p12</td>
</tr>
<tr>
<td>nv15 (351aa)</td>
<td>ND</td>
<td>62.5% cytoplasmic</td>
<td>Nucleus, perinuclear region</td>
<td>XP_358010.r (97%)</td>
<td>NP_959176.m (67%)</td>
<td>NP_900725.m (74%)</td>
<td>20p12.1-16p12</td>
</tr>
<tr>
<td>nv16 (60aa)</td>
<td>ND</td>
<td>60.9% nuclear, 21.7% cytoplasmic</td>
<td>Cytoplasm, focal adhesions</td>
<td>XP_250734.r (92%)</td>
<td>XP_541912.r (78%)</td>
<td>XP_418803.m (83%)</td>
<td>XP_691259.d (60%)</td>
</tr>
<tr>
<td>nv17 (351aa)</td>
<td>ND</td>
<td>62.5% cytoplasmic</td>
<td>Nucleus, perinuclear region</td>
<td>XP_358010.r (97%)</td>
<td>NP_959176.m (67%)</td>
<td>NP_900725.m (74%)</td>
<td>20p12.1-16p12</td>
</tr>
<tr>
<td>nv18 (361aa)</td>
<td>ND</td>
<td>62.5% cytoplasmic</td>
<td>Nucleus, perinuclear region</td>
<td>XP_358010.r (97%)</td>
<td>NP_959176.m (67%)</td>
<td>NP_900725.m (74%)</td>
<td>20p12.1-16p12</td>
</tr>
</tbody>
</table>

**Abbreviations:** aa, amino acid; Ac., accession; at, Arabidopsis thaliana; ce, Caenorhabditis elegans (nematode worm); cf, Canis familiaris (dog), dr, Danio rerio (zebrafish); dm, Drosophila melanogaster (fruitfly); gg, Gallus gallus (chick); NA, not analysed; ND, none detected; NLS, nuclear localisation signal; gn, Rattus norvegicus (rat); sc, Saccharomyces cerevisiae (budding yeast); tn, Tetraodon nigroviridus (pufferfish); xl, Xenopus laevis (frog).
Figure 2. Subcellular distributions of novel uncharacterised molecules selected for further investigation from the developmentally enriched cDNA library: nv1 (A, B), nv2 (C, D), nv4 (E, F), nv5 (G, H), nv6 (I, J), nv11 (K, L), Trp53inp2 (M, N). Cultured BHK-21 and HeLa cells were transiently transfected with mouse HA or myc epitope-tagged full-length predicted coding sequence constructs for each molecule investigated. For BHK-21 transfections, DAPI stained and merged images are shown alongside images of transfected cells detected with monoclonal anti-HA or anti-myc antibody. nv1-HA appears to localise to endoplasmic reticulum (A, B), nv2-HA (C, D) and nv6-myc (I, J) show nuclear and cytoplasmic distribution, nv4-HA (E, F) and nv5-HA (G, H) appear to be predominantly nuclear, nv11-myc appears to be localised to the nucleus and mitochondria (K, L), and Trp53inp2 appears to be distributed to the nucleus and in puncta in the perinuclear region (M, N). All images are resized to equivalent scale, scale bar in (A) equals 10 microns.
Figure 3. Subcellular localisations of poorly characterised molecules selected for further investigation from the developmentally enriched cDNA library: Hn1 (A, B), Fatso (C, D), Elp3 (E, F), Limd1 (G, H), Thoc3 (I, J), Maged2 (K, L), Psf2 (M, N). Cultured BHK-21 and HeLa cells were transiently transfected with mouse HA or myc epitope-tagged full-length predicted coding sequence constructs for each molecule investigated. For BHK-21 transfections, DAPI stained and merged images are shown alongside images of transfected cells detected with monoclonal anti-HA or anti-myc antibody. Hn1 (A, B) and Psf2 (M, N) show nuclear and cytoplasmic distribution, Fatso is predominantly nuclear (C, D), Elp3 appears to have a nuclear and mitochondrial localisation (E, F), Limd1 appears to be localised across the cytoplasm and at focal adhesions (G, H), Thoc3 exhibits aggresomal and vesicle-like distribution (I, J), and Maged2 is localised to the nucleus (concentrated in nucleoli) and cytoplasmic endosomal-like structures (K, L), and also appears to be membrane-associated (L). All images are resized to equivalent scale, scale bar in (A) equals 10 microns.
Figure 4. Co-localisations of a subset of transiently transfected novel/uncharacterised molecules with known markers for subcellular compartments and structures. Nv1-HA (A) colocalises with transfected I-CAT-GFP (B) within the endoplasmic reticulum (A-C). Nv11-myc (D) and Elp3-HA (G) colocalise with the mitochondrial marker MitoTracker Red (D-I). Limd1-HA (J) colocalises with vinculin (K) at focal adhesions (J-L). Arrows indicate observed colocalisation. Scale bars are equal to 10 microns.
Eleven of the 17 genes selected for further analysis represent molecules which are either novel or had no previously characterised functional role (nv1/NM_134022, nv2/AK081685, nv4/NM_026455, nv5/AK013921, nv6/NM_025607, nv9/AK141733, nv11/NM_133924, Trp53inp2, Zcchc3, Hn1, Fio). For one of these (nv9/AK141733) a partial protein is predicted, however sequence data was insufficient to determine a putative full-length transcript and corresponding protein. Five of the 17 candidates had published functional roles but had not been previously described in embryonic development (Elp3, Limd1, Thoc3, Maged2, Psf2). One molecule (nv13) was identified for which originally there were no corresponding proteins identified in the NCBI public database, although our embryonic expression data (Figure 1Q) suggested that it was likely to represent an expressed RNA transcript. The most recent mouse genome annotation (FANTOM III) suggests that nv13 may represent a potential protein-coding transcript AJ290943 (although this is not confidently predicted), but also identifies it as a complementary antisense sequence for the mRNA transcript of RhdB7 (Rhomboid, veinlet-like 7) (23). RhdB7 is a novel predicted transmembrane protein with similarity to Drosophila Rhomboid protein, a known integral membrane serine protease important in dorso-ventral axis establishment and development of the nervous system (24-26). Non-coding RNA transcripts could potentially play a role in regulation and modulation of expression of their complementary genes, thereby having a relatively unique and unexplored functional role in embryogenesis (reviewed in (27)). Alternatively this novel transcript could represent an expressed pseudogene, and this still remains to be determined.

5.2. Prediction of domains, subcellular localisation and transfection of tagged constructs

While many key developmental genes are known to be involved in signal transduction and transcriptional regulation, there is increasing evidence that other classes of molecules, such as those involved in cellular trafficking, play important roles in regulating the processes of embryogenesis (1, 28). Analysis of subcellular localisation of uncharacterised proteins has the potential to shed light on their function at the cellular level, and thereby provide insight into the role of these proteins in the developing embryo. For the 17 candidates selected, prediction tools were utilised to analyse amino acid sequences for known or predicted mouse proteins (summarised in Table 2). Additionally, tagged constructs for 14 protein-coding genes were generated and transiently transfected into mammalian cell lines to reveal subcellular localisation (Figure 2, Figure 3). Nine of the 14 proteins analysed showed a subcellular localisation which was consistent with that predicted by PsortII (nv2/NP_766535, nv4/NP_080731, nv6/NP_079883, nv11/NP_598695, Trp53inp2, Hn1, Fio, Maged2 and Psf2) and three of the remaining five had localisations that were consistent with the known functions of their predicted domains (nv5/NP_001001737, Elp3 and Limd1; Table 2). This correlation between predicted and experimental localisations supports the value of the utilisation of prediction tools in analysing protein sequences to glean informative data. For the remaining two proteins, predicted localisations did not agree with experimental localisations observed for over-expressed proteins (nv1/NP_598783 and Thoc3). This could be due to the fact that the myc/HA epitope tag may cause the protein to misfold and/or mislocalise, or because the cell type used was not physiologically relevant and correct localisation required the presence of interacting partners. Alternatively, the prediction databases used may not detect domains or motifs which have not yet been well characterised or for which amino acid identity of the query sequence is below threshold level.

Of the molecules analysed, four contain domains typical of and/or are known to be involved in transcriptional regulation, including nv5/NP_001001737 (Kruppel-associated box), nv6/NP_079883 (C3HC4 RING zinc finger), Elp3 (histone acetyltransferase), and Limd1 (Lim domains, known to repress E2F-driven transcription (29)). Transcriptional activation and repression via molecules such as transcription factors and chromatin modifiers is known to be a pivotal mechanism for coordination of pathways involved in embryonic development. Also, two proteins contain motifs typical of molecules involved in protein-protein interactions, nv11/NP_598695 (coiled-coils) and Thoc3 (WD40 repeats).

Concurrent with our study, Limd1 was described as a tumor suppressor which binds the retinoblastoma protein (pRB) and represses E2F-driven transcription (29). In the same study, it was also shown to shuttle between the nucleus and cytosol in U2OS cells (human osteosarcoma), and our study showed that in addition, this protein is localised to focal adhesions in BHK-21 and HeLa cells. This is consistent with the observed localisations of members of the same protein family, Ajuba, Zyxin, Lipoma Preferred Partner, Migfilin, Trip6 and Zrp1 (30-35), and with the findings of a very recent study published while our manuscript was in preparation, which also showed localisation of LimD1 to focal adhesions (36). Furthermore, a number of these related proteins have also been shown to shuttle between the nucleus, cytosol and focal adhesions, in a manner dependent on their interactions with various other molecules (reviewed in (37)). Focal adhesions function in cell migration, cell-substratum adhesion and intracellular signalling, processes essential to embryogenesis as well as a range of pathological processes including tumorigenesis and metastasis (reviewed in (38)).

5.3. Potential association of novel and poorly characterised molecules with human hereditary disorders

A focus of this study is the potential association of novel genes with human developmental disorders. To facilitate this, human orthologues for genes which showed restricted embryonic expression were mapped to human chromosomal loci using in silico techniques. Databases were searched to determine any disorders previously mapped to the same chromosomal regions which were characterised by appropriate phenotypes based primarily on sites of expression in mouse embryos. Of the 17 that were
investigated, ten genes were shown to be mapped at or nearby chromosomal regions to which relevant human disorders have also been linked. These genes represent putative candidate genes responsible for these disorders, and a number of candidates are currently under investigation to search for potential disease-causing mutations. In other cases, further fine mapping of disease loci relative to these genes is required to assess the likelihood of their involvement.

6. CONCLUSIONS AND SUMMARY

Based on large-scale screening of a subtracted developmentally enriched pharyngeal arch library, we have identified a number of novel molecules expressed in specific regions of the developing mouse embryo, as well as molecules which had not been previously associated with developmental processes. A number of these molecules represent positional candidates for known developmental disorders based on their chromosomal localisation and sites of expression in the developing embryo. The use of a predictive in silico approach in combination with techniques to investigate gene expression alongside subcellular localisation provides a powerful tool to prioritise molecules for further characterisation, and can also hint at potential functional roles which are worthy of future investigation.

7. ACKNOWLEDGMENTS

The authors would like to thank Vicki Metzis for technical assistance and staff of the Physiology and Pharmacology SPF and Queensland Biosciences Precinct (The University of Queensland) animal houses for help with mouse husbandry and breeding. JSB is a recipient of Australian Postgraduate Award, FS is a UQ Postdoctoral Fellow, CW is an Australian National Health and Medical Research Council (NHMRC) Senior Research Fellow, and RDT is supported by an NHMRC R. Douglas Wright Career Development Award. This work was supported by an NHMRC project grant (CW). The IMB incorporates the Centre for Functional and Applied Genomics, a Special Research Centre of the Australian Research Council. Confocal microscopy was performed at the ACRF/IMB Dynamic Imaging Facility for Cancer Biology, established with funding from the Australian Cancer Research Foundation.

8. REFERENCES


Novel genes in the mouse embryonic facial primordia


Novel genes in the mouse embryonic facial primordia


**Abbreviations:** BLAST, basic local alignment search tool; DIG, digoxigenin; E10.5, embryonic day 10.5; GFP, green fluorescent protein; HA, hemagglutinin; nv, novel molecule; ORF, open reading frame; PCR, polymerase chain reaction; UTR, untranslated region.

**Key Words:** Mouse, Gene, Craniofacial Development, Pharyngeal Arches, Novel Genes, *In Situ* Hybridisation, Subcellular Localisation

**Send correspondence to:** Dr Carol Wicking, Institute for Molecular Bioscience, The University of Queensland, St Lucia 4072, Queensland, Australia. Tel: 617-3346-2052, Fax: 617-3346-2101, E-mail: c.wicking@imb.uq.edu.au

http://www.bioscience.org/current/vol11.htm