Chapter 2

Materials and Methods

2.1 Materials

The majority of chemical reagents were bought from Sigma Chemical Co., BDH Chemical Ltd., and Difco Laboratories unless stated in the text. Similarly restriction enzymes were bought from New England Biolabs unless stated differently. The sources of the commercial kits used in this thesis purchased from various companies are stated in the text. All primers were synthesised in-house by the Sanger Institute Oligo Preparation laboratory. PCR was generally performed using Amplitaq® DNA Polymerase from Perkin Elmer.

2.1.1 Solutions, buffers and media

Solutions used in this thesis are listed according to the methods section they were used. All solutions were made up in double-distilled water (ddH$_2$0), unless stated otherwise.

used in section 2.4

2 x TY: 15 mg/ml bacto-tryptone, 10 mg/ml yeast extract, 5 mg/ml NaCl (pH 7.4) ddH$_2$0 up to 1 litre and autoclave to sterilise.

Chloramphenicol: 30 mg/ml stock made up in 100% ethanol, filtered to sterilise and stored at -20°C. Used at 30µg/ml final concentration.

GTE: 50 mM Glucose, 25 mM Tris (pH7.5), 10 mM EDTA

NaOH/SDS: 0.2M NaOH, 1% (w/v) SDS

3 M KOAc (pH5.5): 300mM potassium acetate (pH4.8), 11.5ml glacial acetic acid, 28.5ml H_20

Boehringer buffer B: 50mM NaCl, 10mM Tris-HCl, 10mM $MgCl_2$, 1mM DTE, pH7.5.

6 x Buffer II: 0.25% bromophenol blue, 0.25% xylene cyanol, 15% ficoll

Vistra green solution: mix 5ml 1 M Tris HCl, 0.5ml 0.1 M EDTA, 50µl Vistra Green (Amersham Life Sciences) made up to 500 ml with ddH_20.

used in section 2.5

10x nick translation buffer: 0.5M Tris-HCl (pH7.5), 0.1M $MgSO_4$, 1mM dithiothreitol, 500µg/ml bovine serum albumin

DNase I (1µg/ml): dilute 10 mg/ml Deoxyribonuclease I (Sigma) stock to 1µg/ml working solution with enzyme diluent.

Enzyme diluent: 500µl glycerol, 100µl nick translation buffer, 400µl ddH_20

Fixative: 3:1 methanol/glacial acid

Formaldehyde fixative: 1% v/v formaldehyde (from 40% stock), 50 mM $MgCl_2$ in PBS. For 50 ml add 1.25 ml formaldehyde, 2.5 ml 1M $MgCl_2$ and make-up to 50mls with 2xSSC.

FISH Hybridisation buffer: 50% deionised formamide, 2x SSC (pH7.0), 10% dextran sulphate, 0.1% SDS, 1x Denhardt's solution, 40mM sodium phosphate pH7.0.

4 x TNFM: 4 x SSC, 0.05% Tween 20, 5 % non-fat milk powder, filtered through several layers of Whatman No.4 filter paper

used in section 2.6

Mung bean nuclease buffer: 100µl 3 M sodium acetate, 250 µl 2 M sodium chloride, 10 µl 1 M zinc chloride, 140 µl water, 500 µl mung bean nuclease (Pharmacia), 500 µl glycerol

Buffered phenol: 1 ml phenol, 200 µl 1 M Tris-hydrogen chloride (shaken and placed on ice for 5 minutes, centrifuged, top layer removed and discarded, 200 µl TE added, mixed, shaken and centrifuged. Kept on ice until needed.

SOC: SOB + 200 µl 20% glucose

SOB: 20 g tryptone, 5 g yeast extract, 10 ml 1M sodium chloride, 0.5 g potassium chloride, water added up to 1 litre

TYE agar: 8 g tryprone, 5 g yeast extract,8 g sodium chloride, 12 g agar, water upto 1 litre

TYE/Amp plates: 2 ml of 25 mg/ml ampicillin was added to 1 ml TYE autoclaved solution which was allowed to cool to 48°C before addition.

IPTG: 40 mg/ml in DMSO. Sterilised by filtration and stored at -20°C.

Xgal: 50 mg/ml in ddH_20. Sterilised by filtration and stored at -20°C.

0.1% DMSO: Dimethyl sulfoxide diluted in ddH_20 and autoclaved.

used in section 2.7

2 x LB: 10 mg/ml bacto-tryptone, 5 mg/ml yeast extract, 10 mg/ml NaCl (pH 7.4) ddH_20 up to 1 litre and autoclaved to sterilise.

Ampicillin: 25 mg/ml Ampicillin stock made up in ddH_20, filtered to sterilise and stored at -20°C.

GET: 30 mM Glucose, 15 mM Tris-HCl (pH8.0), 30 mM Na_2EDTA, 60 µg/ml RNase A

Bind solution: (6.1 M Potassium Iodide) 40 g potassium iodide in 28 ml ddH_20. Stored in the dark at room temperature.

Precipitation mix: 100 ml 96% ethanol, 2 ml 3 M sodium acetate, 4 ml 0.1 mM EDTA

Sequencer Loading dye: 25 mM EDTA (pH 8.0), 50 mg/ml Blue dextran, deionised formamide (5:1 formamide: EDTA/Blue dextran).

used in sections 2.8-2.13

10 x PCR buffer: 500 mM KCl, 50 mM Tris (pH8.5), 25 mM $MgCl_2$.

PBS: 10 g sodium chloride, 0.25 g potassium chloride, 1.44 g sodium hydrogen phosphate (dibasic), 0.25 g potassium dihydrogen phosphate, made up to 1 litre with water and made to pH with sodium hydroxide. Stored at 4°C.

4 x Spotting buffer: 1 M sodium phosphate buffer pH 8.5, 0.001% sarkosyl

Bacterial mRNA "cocktail": pool of cDNA bacterial clones for B.subtilis *trp* gene (30 ng/µl), *lysA* gene (0.3 ng/µl), *thrB* gene (3 ng/µl), *dapB* gene (15 ng/µl), *pheB* gene (1.5 ng/µl) all purchased from American Type Culture Collection (ATCC catalogue numbers 87482 to 87486)

Microarray hybridisation buffer: 5 x SSC, 6 x Denhardts solution, 60mM Tris HCl (pH7.6), 0.12% sarkosyl, 48% formamide filter sterilised

100 x Denhardt's solution: 20 mg/ml Ficoll 400-DL, 20mg/ml polyvinylpyrrolidine 40, 20mg/ml BSA (pentax fraction V)

Microarray wash solution 1: 2 x SSC, filter sterilised

Microarray wash solution 2: 0.1 x SSC, 0.1 % SDS, filter sterilised

Microarray wash solution 3: 0.1 x SSC, filter sterilised

used in section 2.14

Spermidine stock: 25.46 mg/ml spermidine (Sigma) in 10 mM Tris (pH 7.4)

1 x denaturation solution: 0.5 M NaOH, 1,5 M NaCl

1 x neutralisation solution: 1.5 M NaCl, 1 M Tris-HCl (pH7.4)

MTN/Southern Wash solution I: 2 x SSC, 0.05% SDS, filter sterilised

MTE Wash solution I: 2 x SSC, 1% SDS, filter sterilised

MTN/Southern Wash solution II: 0.1 x SSC, 0.1% SDS, filter sterilised

MTE Wash solution II: 0.1 x SSC, 0.5% SDS, filter sterilised

General solutions, buffers and media used in this thesis

10 x TBE: 890 mM Tris base, 890 mM Borate, 20 mM EDTA (pH8.0)

50 x TAE: 2 M Tris base, 5.7% v/v glacial acetic acid, 50 mM EDTA

20 x SSC: 175.3 g sodium chloride, 88.2 g sodium citrate, made up to 1 litre with ddH_20.

1 x $T_{0.1}E$: 10 mM Tris-HCl (pH8.0), 0.1 mM EDTA

1 x TE: 2 ml Tris (pH7.4), 200 µl 0.1 M EDTA, ddH_20 to 200 ml.

10 mM dNTPs mix: 1 ml of each dNTP (100 mM) in 6 ml of ddH_20. Stored at -20 °C.

1µl 10 mM dA,T,GTP/5 mM dCTP mix: 25 µl dATP, 25 µl dTTP, 25 µl dGTP, 10 µl dCTP and 15 µl ddH_20.

2.1.2 Loading dyes

Loading dye: 5 mg bromophenol blue, 0.5 g Ficoll, 0.5 ml 10 x TBE, 4.5 ml ddH_20

Loading buffer: 10 µl 10 x TBE, 20 µl loading dye, 50 µl water

Sequencer loading dye: 25 mM EDTA (pH 8.0), 50 mg/ml Blue dextran, deionised formamide (5:1 formamide: EDTA/Blue dextran).

2.1.3 Nucleotides

Amersham Biosciences	Redivue™ deoxycytidine 5'-[α-^{32}P]-dCTP-triphosphate, triethylammonium salt (AA0075)
Invitrogen	Renaissance R Cyanine 3-dCTP (NEL576) Renaissance R Cyanine 5-dCTP (NEL575)
Pharmacia	100mM dATP, dCTP, dGTP, dTTP (27-2035-01)
Boehringer	Biotin-16-dUTP (1 mM) (1093-070)

2.1.4 Size markers and ladders

1kb DNA ladder (1 µg/µl) (Gibco BRL Life Technologies): 5 µl 1 Kb DNA ladder mixed with 1 µl 50 x TAE, 10 µl Ficoll dye and 34 µl ddH$_2$0. The 1 Kb DNA ladder contains 1 to 12 repeats of a 1018 bp fragment and vector fragments from 75 to 1636 bp to produce the following sized fragments in bp: 75, 142, 154, 200, 220, 298, 344, 394, 516/506, 1018, 1635, 2036, 3054, 4072, 5090, 6108, 7125, 8144, 9162, 10180, 11198, 12216.

100 bp DNA ladder (1µg/µl) (Gibco BRL Life Technologies): 50µl (1µg/µl) 100 bp

DNA ladder, 60 µl loading buffer and 390 µl ddH_20. The 100 bp DNA ladder consists of 15 blunt-ended fragments between 100 and 1500 bp in multiples of 100 bp and an additional fragment at 2072 bp. The 600 bp fragment is approximately 2 to 3 times brighter that the other ladder bands to provide orientation.

Fingerprinting gel marker: 19.2µl $T_{0.1}E$, 1.5µl Analytical Marker DNA wide range (Promega), 0.2µl Molecular Weight Marker V (Boehringer-Mannheim) and 4.2µl 6x loading dye were added to a 1.5ml microfuge tube and stored at -20°C). The Analytical Marker DNA wide range provides an evenly spaced distribution of DNA fragments from 0.702 kb to 29.95 kb.

Lambda DNA-Hind III marker (Gibco BRL Life Technologies): 8 µl lambda DNA-*Hind* III digest, 60 µl TBE buffer, 252 µl ddH_20 was incubated at 65°C for 5 minutes then snap chilled on ice before 80 µl loading dye was added. The *Hind* III digest of lambda DNA yields 8 fragments suitable for use as molecular weight standards for agarose electrophoresis of the following sizes; 125, 564, 2027, 2322, 4361, 6557, 9416, 23130.

Lambda Hind III/pBR322 marker: 8 µl lambda *Hind* III (NEB), 60 µl 10 x TAE, 252 µl ddH_20 were heated at 65°C for 5 minutes then snap chilled on ice then 6 µl pBR322 BstNI (NEB) and 80 µl loading dye were added.

2.1.5 Sources of DNA and RNA

Human genomic DNA was purchased from Clontech (catalogue number 6550-1). Total RNA was extracted from Raji, Jurkat, 293T and U937 cell-lines (a kind gift from John Trowsdale, Division of Immunology, Department of Pathology, University of Cambridge) and a THP1 cell-line (kindly provided by Paul Lehner, CIMR, Cambridge). Total human Adrenal Gland, Brain, Skeletal Muscle, Spleen and Testis RNA were purchased from Ambion (catalogue numbers 7994, 7962, 7982, 7970 and 7972, respectively). Universal Human Reference RNA was purchased from Stratagene (catalogue number 740000). The human multiple tissue expression (MTE™) array and the multiple tissue northern (MTN™) were purchased from Clontech (catalogue numbers 7776-1 and 7760-1, respectively). Information regarding the sources of RNA can be found on the web-site http://www.clontech.com.

Methods

2.2 Agarose gel preparation and electrophoresis

Unless stated otherwise in the text: agarose gels were prepared in either 1x TBE or 1 x TAE buffer containing 250ng/µl ethidium bromide and the appropriate percentage of agarose was used according to the size of fragments being separated; a 2.5% agarose gel was used for electrophoresis of fragments below 1kb, and a 0.8-1.0% agarose gel for analysis of larger fragments. Electrophoresis was performed at 50-100 V for 15-45 minutes depending on separation required. The sizes of the DNA fragments were estimated by running either the 1 kb or 100 bp ladder size standards.

2.3 Sequencing gel

The denaturing acrylamide gel (6%) was made up using 30 g urea in 9 ml acrylamide/bisacrylamide solution, 4 ml 10 x TBE and 37 ml ddH_20. The urea was dissolved by heating to 60°C and stirring. The solution was made up to 60 ml with water and placed in a dessicator for 4 minutes. Just prior to pouring the gel, 138 µl of 25% ammonium persulphate and 138 µl TEMED were added. The gel was then carefully syringed between the glass plates whilst tapping the glass gently to get rid of air bubbles. The appropriate comb was inserted and the gel was left to set for at least 90 minutes prior to use.

Mapping and sequencing

2.4 Restriction Digest Fingerprinting

The BAC genomic clone, bA465F21 (AC006313) was fingerprinted using the *HindIII* digest fingerprinting method essentially as described by Olson *et al*, 1986.

2.4.1 Filterprep isolation of BAC DNA

1. 500 µl of 2 x TY containing 30 µg/ml of chloramphenicol were added to a 96-well 1 ml Beckman box.
2. Each well was inoculated from a glycerol stock with either a 96-well inoculating tool, or a sterile cocktail stick. A plate sealer was placed on top of a plate to seal the wells and the cultures grown for 16-18 hours at 37°C with gentle shaking (300 rpm).
3. For each well, 250 µl of the overnight growth were transferred to a clean round-bottomed Corning microtitre plate using a 50- to 250-multichannel pipette (Finnpipette). The cells were pelleted by centrifugation at 2500 rpm at 20°C from 4 minutes.
4. For each well, the supernatant was discarded and the pellet re-suspended in 25 µl of GTE and mixed gently by vortexing. 25 µl of freshly prepared NaOH/SDS solution was added and mixed by tapping the plate gently and left to stand at room temperature for 5 minutes.
5. 25 µl of chilled 3 M KOAc (pH5.5) solution were added, mixed and left at room temperature for 5 minutes.
6. A microtitre plate containing 100 µl of isopropanol was taped to the bottom of

2 µm filter-bottomed plate (Millipore). The total well volume of the sample was transferred to the filter-bottomed plate.

7. These 2 plates were then spun at 2500 rpm, 20°C for 2 minutes to ensure all liquid had been transferred from the filter plate to the lower plate; the filter plate was then discarded.
8. After separation from the filter plate, the lower (Corning) plate was left at room temperature for 30 minutes before being centrifuged at 3200 rpm, 20°C for 20 minutes.
9. The supernatant was discarded from the plate and the DNA pellet was briefly dried by inverting the plate and placing on clean tissue paper.
10. 100 µl of 70% ethanol were added to the dried DNA to wash the pellet, mixed gently, and the DNA precipitated by centrifuging at 3200 rpm 20°C for 10 minutes. This step was repeated.
11. Finally, the supernatant was discarded and the DNA pellet was dried before being resuspended in 5 µl of fresh $T_{0.1}E$ with RNase (1µg/ml).
12. Samples were stored at -20°C.

2.4.2 Restriction digest fingerprinting (*Hind* III) of BAC DNA

1. For one 96-well microtitre plate of sample DNA, a premix containing 286 µl ddH_20, 99 µl Boehringer buffer B, 55 µl *Hind* III was prepared in a 1.5 ml microfuge tube, and mixed by vortexing.
2. 4 µl of the premix was added to each well of a 96 well-microtitre plate containing previously prepared DNA (see section 2.4.1) and mixed gently by vortexing at 37°C for 2 hours.

3. The reaction was terminated by adding 2 µl of 6x Buffer II and either stored at 4°C or loaded immediately.
4. 0.8 µl of the fingerprinting marker was added to the first well and then every sixth well of a freshly prepared 1% agarose/1 x TAE gel. 1 µl of each sample was loaded (i.e. wells 2-5, 7-10 *etc)* between the marker lanes. Fragments were resolved by electrophoresis through the gel at 4°C in a cold room for 15 hours at 90 volts.
5. Following electrophoresis, the gel was cut down so the length was 19-20 cm and stained with Vistra Green solution for 30-45 minutes on a shaker. The gel was washed with ddH_20 to remove excessive stain.
6. The gels were scanned on a FluorImager SI. The parameters were set to 530 nm for emission filter, the pixel size was 100 microns, detection sensitivity was normal, digital resolution was at 16 bits, dye was single label, excitation filter was 488 nm, Em filter 1530 nm and PMT voltage was 800.
7. The gel image was transferred to a UNIX workstation and entered into the fingerprint 'IMAGE' analysis system (Sanger Institute in house software).The band pattern was extracted using 'IMAGE' and the data entered into another program, fingerprinted contigs, FPC (Soderlund *et al*, 1997), where the fingerprint patterns were compared to those already in the database and the position of the clone within a contig determined.

2.5 Fluorescent *in-situ* hybridisation (FISH) mapping

Cytogenetic mapping using FISH techniques were performed using chromosome 9 clones. The BAC clone, bA465F21, was fluorescently labelled and hybridised to metaphase chromosomes to determine which chromosome it maps to (Pinkel *et al*, 1986). In addition, the orientation and order of 3 contigs were resolved by interphase FISH (Wilke *et al*, 1994) and the sizes of gaps between 5 contigs determined using extended DNA fibres using Fibre FISH (Heiskanen *et al*, 1994).

2.5.1 Labelling of FISH probe using Nick translation

1. 1 µg of clone DNA was labelled with 1mM biotin-16-dUTP (Boehringer) in a 25 µl reaction containing; 2.5 µl nick translation buffer, 1.9µl 0.5 mM dATP, dCTP and dGTP mix, 0.7 µl 1mM biotin-16-dUTP, 1µl DNase I* (Sigma), 0.5 µl DNA polymerase I (10U/µl Sigma) made-up to 25 µl with H_20.

*In order to determine the concentration of DNase I and incubation time a series of dilutions were carried out using different amounts of DNase I in 50 µl reaction volumes containing; 2 µg test DNA, 5 µl nick translation buffer, 1-2 µl DNase I (1µg/ml working stock in enzyme diluent). The reactions were incubated at 14°C for 60 minutes. A 10 µl aliquot was removed after 20 minutes with further 10 µl aliquots removed at 10 minute intervals. All samples were run on a 1% agarose gel and the DNase I concentration and incubation time which gave fragment smears with a size range of 200-700 bp used.

2. The 25 µl reaction was incubated at 14°C for 60 minutes and the labelling reaction terminated by adding 2.5 µl of 0.5 M EDTA (pH8.0)
3. 2.5 µl 3M sodium acetate (pH7.0) and 60 µl of 100% ethanol were added to the reaction and the probe precipitated at -70°C for 30 minutes.
4. The mixture was centrifuged at 13,000rpm for 10 minutes and the pellet washed twice with the addition of 500 µl 70% ethanol and centrifuged at 13,000 rpm for a further 2 minutes. The pellet was air-dried at 37°C.

5. The pellet was resuspended in 10 µl $T_{0.1}E$ and 2 µl of sample was run on a 1% agarose gel to check efficacy of the reaction.

2.5.2 Preparation of microscope slides

The slides containing the extended DNA fibres and the metaphase and interphase cell-suspensions were kindly provided by the Sanger Institute Molecular Cytogenetics group.

1. Microscope slides were washed in 2% Decon and sonicated in a sonicator bath then rinsed under cold running water for 60 minutes. Stored in 96% ethanol.
2. The slides were removed from the ethanol and polished with a dry, lint-free tissue.
3. The metaphase or interphase cell-suspension was mixed by gentle flicking of the tube and a single drop was dropped onto the slide using a Pasteur pipette.
4. A drop of fixative was added whilst the first drop was still spreading and the slide air-dried.
5. The slides were fixed in a coplin jar of fixative at room temperature for 30-60 minutes, air-dried and stored in a sealed box at room temperature until needed.
6. Prior to use the slides were incubated in 2 x SSC at 37°C for 5 minutes, followed by 5 minute incubation at 37°C in 0.01 M HCl and 10 µl of 25% pepsin (in ddH_20).
7. The slides were rinsed 3 times in 2 x SSC for 2 minutes each at room temperature and then fixed in formaldehyde fixative for 10 minutes also at room temperature.
8. The slides were rinsed again 3 times in 2 x SSC for 2 minutes at room

temperature then dehydrated through exposure to 3 concentrations of ethanol: slides were incubated at room temperature with 70%, 70%, 90%, 90%, 100% ethanol for 1 minute each.

9. The slides were air dried to evaporate the remaining ethanol.

2.5.3 Hybridisation of FISH probes

1. 1 µl Cot1 DNA (1µg) and 11.5 µl FISH hybridisation buffer was added to 0.5 µl labelled DNA (30-50 ng), mixed thoroughly and denatured at 65°C for 10 minutes.
2. The denatured probe was then incubated at 37°C for 1 hour.
3. Prior to hybridisation the slides were denatured in 70% formamide (in 2 x SSC) at 65°C for 2 minutes then quenched in ice cold 70% ethanol then dehydrated through an ethanol series (70%, 70%, 90%, 90%, 100% for 1 minute each) and air-dried.
4. The hybridisation mix was placed onto the denatured slides and covered with 22 x 22 mm cover slip. Rubber cement was used to seal the cover slips and the slides were incubated in a moist chamber at 37°C for 24 hours.
5. After hybridisation the cover slip was removed and the slides rinsed in 2 x SSC for 5-20 minutes.
6. The slides were washed twice at 42°C in 50% formamide (in 2 x SSC) for 5 minutes.
7. Further washing was performed at 42°C in 2 x SSC for 5 minutes.
8. For detection of biotinylated probes 100 µl of 4 µg/ml avidin Texas Red DCS (Vector) was added and the slide covered with Nescofilm and incubated at

37°C for 20-60 minutes in a humid chamber.

9. The slides were washed in 4 x TNFM at 42°C for 5 minutes.
10. The slides were drained and 100 µl of 4 µg/ml biotinylated anti-avidin D plus 1:500 dilution of mouse anti-digoxin (Sigma) was added. The slide covered in Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
11. The slides were washed in 4 x TNFM at 42°C for 5 minutes.
12. The slide covered in Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
13. The slides were drained and 100 µl of 4 µg/ml avidin Texas Red DCS plus 10 µg/ml goat anti-mouse FITC conjugate (Sigma) was added. The slide covered in Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
14. The slides were washed in 4 x TNFM at 42°C for 5 minutes.
15. The slides were drained and 100 µl of 4 µg/ml avidin Texas Red DCS (Vector) was added and the slide covered with Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
16. The slides were washed in 4 x SSC, 0.05% Tween 20 at room temperature and counterstained in 0.08 µg/ml DAPI (4',6'-diamidino-2-phenylindole hydrochloride) in 2 x SSC for 2-3 minutes.
17. The slides were rinsed in 2 x SSC then dehydrated through the ethanol series (70%, 70%, 90%, 90%, 100% for 1 minute each) and air-dried.
18. 20 µl of antifade solution (Citifluor AF1) was added to a clean 22 x 32 mm cover slip and overlayed on the slide. The cover slip was sealed using nail varnish.
19. The slides were analysed using a Zeiss Axioscop fluorescence microscope equipped with a CCD camera. Separate images of the DAPI staining of the

chromosomes and the biotinylated probes were merged using SmartCapture software (Digital Scientific Ltd).

2.6 Production of shotgun libraries for shotgun sequencing (essentially as described by Bankier *et al*, 1987)

The minimum set of clones to cover chromosome 9 were selected for sequencing using the large-scale maps produced by FPC fingerprinting methods. Each clone is divided into fragments by sonication which are then assembled so overlapping fragments of sequence provide the complete sequence across the clone. The random nature of sonication produces fragments that will be sequenced on average 6-8 times before a project is considered complete; this redundancy is necessary to ensure that sequencing errors are resolved. The chromosome 9 BAC clone DNA for bA18B16 was provided by the Sanger Institute Sub-cloning laboratory and sub-cloned by me.

2.6.1 Sonication and subfragment end repair of plasmid DNA

1. In order to estimate the concentration of DNA of the BAC clone, a 0.5 % agarose, 1 x TBE gel was run on a 10 x dilution of the BAC. The DNA sample was diluted 1/10 in $T_{0.1}E$ and 1 µl was run alongside lambda *Hind* III/pBR322 marker. Samples were visualised by soaking the gel in 500 ml of 1 x TBE containing 25 µl ethidium bromide (10 mg/ml) for 5 minutes the de-stained in ddH_20 for 10 minutes.
2. From the gel image, the amount of DNA required to obtain 10 µg was taken for sonication.
3. To the 10 µg DNA ddH_20 was added to a final volume of 54 µl. 6 µl of mung

bean buffer was added, mixed and collected by centrifugation.

4. The sample tube was placed in the 'cup-horn' of the sonicator containing ice cold water 1 mm from the face of the probe.
5. An output of approximately 12% on the 400 watt Virsonic 300 sonicator was used for 10 seconds in order to produce fragments of the required length.
6. 1 µl of sonicated DNA was mixed with 4 µl of loading buffer and the sample was run alongside lambda *Hind* III/pBR322 markers on a 0.8% agarose gel with 1 x TBE.
7. If sonication was successful the DNA was visible as a smear with no sign of a band of high molecular weight DNA. If a band was visible the samples were sonicated for a further 5 seconds and checked again on a 0.8% / 1 x TBE agarose gel.
8. The ends of the sonicated DNA fragments were repaired by adding 0.3 µl of mung bean nuclease buffer to the DNA. This mixture was placed in a 30°C water bath for 10 minutes.
9. The volume in the tube was made up to 200 µl with H_20, 20 µl of 1 M sodium chloride, 550 µl of ice cold 100% ethanol and 1 µl of pellet paint (Novagen) were added to the DNA.
10. In order to precipitate the DNA, it was left for 2-18 hours at -20°C and then centrifuged for 30 minutes at 4°C at 13,000 rpm.
11. The supernatant was removed from the tube and the DNA pellet was washed in 1 ml 100% ethanol by centrifugation for 10 minutes at 4°C at 13,000 rpm.
12. The ethanol was removed and the pellet was dried in a vacuum dryer for 10-15 minutes.

2.6.2 Selection of suitably sized DNA fragments for subcloning

1. The pellet was thoroughly resuspended for loading in 6.25 µl $T_{0.1}E$, 0.75 µl 10 x TAE and 2 µl loading dye.
2. All 9 µl of sample was loaded on a 0.8% agarose/1 x TAE gel with a lambda *Hind* III/pBR322 marker for 2 hours at 35 mA, 50-60 v.
3. The bands were visualised on an ultra violet transilluminator (312 nm) and the bands corresponding to 1.4-2 Kb (ideal) size were cut out. Additional bands of 1-1.4 Kb and 2-4 Kb were also cut from the gel and stored at 4°C. The pieces of gel were weighed to estimate the gel volume.
4. The 1.4-2 Kb gel fragment was placed in a tube and incubated at 65°C for 5-10 minutes.
5. 4 µl of AgarACE (Promega) was added to the tube in a 42°C waterbath. The molten gel was incubated at 42°C for 15 minutes.
6. 15 µl of sodium chloride, 150 µl of buffered phenol and the appropriate volume of $T_{0.1}E$ buffer corresponding to the weight of the gel piece was added to the tube to a final volume of 135 µl.
7. The tube was mixed by vortexing and centrifuged at 13,000 rpm.
8. The upper (aqueous) phase (approximately 230 µl) was extracted and added to the tube containing the 1.4-2 Kb gel fragment. 30 µl of $T_{0.1}E$ was added to the bottom layer, vortexed and centrifuged at 13,000 rpm for 3 minutes.
9. The upper (aqueous) phase was removed and pooled with the first layer removed.
10. 1 µl of pellet paint (Novagen) and 350 µl 100% ethanol were added to the tube which was placed at -20°C overnight.
11. The tube was centrifuged at 4°C at 13,000 rpm for 30 minutes and the ethanol

was discarded.

12. The pellet was resuspended in 1 ml of ethanol and spun at 4°C, 13,000 rpm for 10 minutes.
13. Ethanol was removed from the pellet which was vacuum dried for 5-10 minutes before resuspension in 5 µl of $T_{0.1}E$.
14. To check for successful elution, 0.5 µl DNA with 4.5 µl loading buffer was run on a 0.8%/ 1 x TBE agarose gel with lambda *Hind* III/pBR322 markers.

2.6.3 Ligation into pUC18 vector

1. A premix of pUC18 (SmaI/CIP, Amersham) and buffer (provided with vector), consisting of 0.05 µl of pUC18 per reaction and 0.1 µl of buffer (supplied with the pUC18) was prepared by vortexing and placing the tube on ice.
2. 0.15 µl of the pUC18-buffer mix was dispensed into the 600 µl Sarstedt tube set-up for each reaction.
3. 0.7 µl of DNA was added to each tube. In addition 3 control tubes were set-up with the following: (a) 0.7 µl ddH_20 (b) 0.7 µl ddH_20 and (c) 0.7 µl Φx174/HaeIII (1.4 ng).
4. 5 µl of mineral oil was added to each tube.
5. With the exception of tube (b), 0.15 µl T4 DNA ligase (Pharmacia) was dispensed to each tube, aiming for the 'bubble' under the oil, and the tubes were mixed and centrifuged for a few seconds.
6. Tubes were transferred to a 16°C incubator and left overnight to allow ligation to occur.

7. Tubes were heated to 65°C for 7 minutes before being left at room temperature for 5 minutes and centrifuged briefly.
8. 49 µl of ddH_20 was added to each reaction and tubes were stored at -20°C until transformations were performed.

2.6.4 Transformation of pUC18 vector

1. 1 µl of ligated DNA was aliquoted into 15 ml glass test-tubes and 500 µl of SOC was added to each 1 ml tube.
2. TG-1 cells (Invitrogen, maintained in 10% glycerol and stored at -70°C) were removed from the freezer and 150 µl 10% glycerol was added to each tube of cells which were left on ice.
3. Cells and glycerol were mixed using a P200 Gilson pipette and 40 µl of this mixture was added to the ligated DNA.
4. The cells, glycerol and DNA were aliquoted into a cuvette placed on ice, then electroporated using a Bio Rad Micropulser at 3.1 ms and 1.9 kv.
5. The cuvette was removed from the Micropulser and 400 µl SOC (pre-warmed to 20-30°C) was added to the cuvette: the mixture of SOC, cells and DNA was taken up and ejected into a test-tube.
6. The test-tubes were incubated in a shaker at 30°C for 1 hour with agitation.
7. TYE/Amp plates (90 mm) were placed at room temperature.
8. The test-tubes were removed from the shaker and 50 µl IPTG (40 mg/ml) 50 µl Xgal (50 mg/ml) were added to each tube.
9. 125 µl of the solution was dispensed into one TYE/Amp plate and 250 µl was dispensed onto a second plate.
10. A sterile spreader was used to make the solution cover the plate in an even

manner.

11. Plates were placed in a 37°C incubator overnight and the number of recombinant (white) and non-recombinant (blue) colonies counted.
12. Successful ligations were stored at -20°C.

2.7 Shotgun sequencing

The chromosome 9 clones, bA18B16 and bA544A12, were sequenced using a modified version of the method described by Sanger *et al* (1977b). Essentially, the DNA was sequenced using the dideoxy termination system in which DNA polymerase uses directed primers to extend a DNA strand from a single stranded template. Extension occurs with the addition of deoxynucleotides complementary to the template strand until the dideoxynucleotide that inhibits further extension is incorporated. The latter are labelled with fluorescent dyes and visualised when separated by gel electrophoresis. The biochemistry will produce populations of products specifically terminating at either A, G, C or T.

2.7.1 Vacuum preparation of template DNA in pUC18 vector

1. 1 ml of 2 x LB containing ampicillin was aliquoted into each well of a 96 well Beckman box, and separate (white recombinant) colonies were picked into each of these wells.
2. Boxes were sealed and the lids were pierced before boxes were placed in a 37°C incubator at 320 rpm and left to grow for 20-24 hours.
3. After growth, 100 μl of the cells were removed from each well and added to a 96 well plate (Corning) containing 50 μl 100% glycerol. The plates were

sealed and stored at -70°C.

4. Boxes were spun for 2 minutes at 4000 rpm, the supernatant was discarded and boxes were inverted on several layers of towels for 20 minutes to remove residual culture supernatant.
5. The pellets were resuspended in 120 µl GET using a plate shaker (Luckham V400 Vortexer) completely resuspended.
6. 120 µl NaOH/SDS solution was added, mixed thoroughly then incubated at room temperature for 2-5 minutes.
7. 120 µl 3 M KOAc (pH 5.5) was added and mixed gently.
8. A filter-bottomed plate (FB; Millipore catalogue number MAFBNOB50) was placed in the bottom of the vacuum manifold (Eppendorf). The lysate was removed from the Beckman box and dispensed into a Multiscreen-NA lysate clearing plate (NA; Millipore catalogue number MANANLY50) which was then placed on top of the manifold.
9. The lysate was drawn through the NA plate into the FB plate inside the manifold by applying the vacuum for 3 minutes not exceeding 8 Hg vacuum setting.
10. The NA plate was discarded and 150 µl of Bind Solution added to the FB plate and mixed.
11. The FB plate was placed on the empty manifold and full vacuum (30 Hg) was applied for 1 minute.
12. The plasmid DNA, bound to the FB plate, was washed with ice cold 80 % ethanol and vacuum filtered at full vacuum for 1 minute.
13. The plasmid DNA was washed again with ice cold 80% ethanol and vacuum filtered at full vacuum for 3 minutes.

14. The FB plate was removed from the vacuum manifold and dried thoroughly at 90°C for 10 minutes or 2 hours at room temperature.
15. 50 µl ddH_20 was added to the centre of each well and left to stand for 5 minutes at room temperature.
16. The plasmid DNA was eluted by placing the FB plate on top of a new microtiter plate (AB gene Thermo-fast® 96 well skirted plate; catalogue number AB-0800) and centrifuging for 2-5 minutes at 4000 rpm.
17. The plasmid DNA was checked on a 0.8 % agarose gel made up in 1 x TBE.

2.7.2 The sequencing reaction

1. 2 µl of DNA was added to 8 µl of a mix made up of 1 µl of forward primer (M13F-21F 5'-TGTAAAACGACGGCCAGT-3'; 6 pM/µl) or reverse primer (pUC18R 5'-GCGGATAACAATTTCACACAGGA-3'; 6 pM/µl), 4 µl BigDye™ Terminator Ready Reaction mix (supplied by PE Applied Biosystems) and 3 µl water.
2. The mixture was centrifuged and placed on a PTC-225 Peltier Thermocycler (MJ Research) with the following program: (i) 96°C for 30 seconds (ii) 50°C for 15 seconds (iii) 60°C for 2 minutes 30 seconds, (iv) repeat (i) – (iii) for 25 cycles (v) 4°C until stopped.
3. To each reaction, 10 µl ddH_20 and 50 µl precipitation mix was added.
4. The plate was centrifuged at 4°C, 4000 rpm for 25 minutes, and the ethanol was decanted.
5. 100 µl of ice-cold 70% ethanol was added, and the plate was centrifuged for 2-3 minutes at 4°C, 4000 rpm. This step was repeated.

6. The ethanol was removed and the plate inverted on a tissue and centrifuged at 250 rpm to remove all traces of ethanol. The plate was dried at 90°C for 10 minutes in the dark. Plates were stored at -20°C until loaded onto the sequencer.

2.7.3 Sequencing instrumentation

DNA sequenced by me was loaded on either an ABI PRISM® 373 DNA sequencer or an ABI PRISM® 377 DNA sequencer and generated by the Sanger Sequencing Centre on an ABI PRISM® 3100 DNA analyser (Applied Biosystems).

2.7.3.1 ABI PRISM® 373 DNA sequencer set-up:

1. The sequencing gel plate (see section 2.3 for preparation) was inserted into the ABI cassette of the ABI PRISM® 373 DNA sequencer and secured using clips; ensuring that the gel plates were flat in the cassette.
2. The plates were cleaned using a lint free tissue and the plates-checked by scanning the glass plates. If 4 flat coloured lines appeared in the scan window the upper buffer chamber was put in place and both upper and lower chambers were filled with 1 x TBE buffer before pre-running the machine for 30 minutes. If there were peaks in the trace the plate was removed from the cassette and cleaned before repeating the plate-checking process.
3. 3 µl of sequencer loading dye was added to each sequencing reaction, briefly centrifuged then denatured by heating at 80°C for 10 minutes before loading.
4. The comb was removed from the gel and wells were rinsed using 1x TBE to

ensure that there were no air bubbles before 3 µl of sample was loaded to each well (36 maximum) using a Gilson pipette.

5. Data was collected over a run-time of 8 hours.

2.7.3.2 ABI PRISM® 377 DNA sequencer set-up:

1. The sequencing gel plate (see section 2.3 for preparation) was inserted into the ABI cassette of the ABI PRISM® 377 DNA sequencer and secured using clips; ensuring that the gel plates were flat in the cassette.
2. The plates were cleaned using a lint free tissue and the plates-checked by scanning the glass plates. If 4 flat coloured lines appeared in the scan window the upper buffer chamber and heat plate that clipped in front of the gel plate were put in place. If there were peaks in the trace the plate was removed from the cassette and cleaned before repeating the plate-checking process.
3. The upper and lower buffer chambers were filled with 1 x TBE buffer before pre-running the machine for 30 minutes.
4. 2 µl of loading dye was added to each dried sequencing reaction and the samples were then briefly centrifuged.
5. The comb was removed from the gel and wells were rinsed using 1x TBE to ensure that there were no air bubbles before 2 µl of sample was loaded to each well (48-60) using a Gilson pipette.
6. Data was collected over a run-time of 4 hours.

2.7.4 Data analysis of shotgun sequencing reactions and clone assembly

The data produced from the ABI sequencers was transferred to the UNIX system where a number of Sanger Institute in house software programs have been developed for the analysis of this data. The first procedure involved in analysing an ABI-PRISM® 373 or ABI-PRISM® 377 sequencing gel is to establish the position of each sample on a gel. This lane tracking is automatically performed by the program ‘Gelminder’ (Platt and Mullikin, unpublished) but manual checking and in some cases, repositioning is required. After manual checking of the lane tracking, the individual bases are called by ‘Gelminder’ using the program ‘Phred’. The sequencing data produced by the capillary sequencer ABI-PRISM® 3100 is automatically uploaded into ‘Capminder’ and the bases are identified using the program ‘Phred’.

Data from each sequencing reaction is then passed into the ‘Automated Sequence Preprocessor (ASP)’ program (Hodgson, unpublished) which cuts off sequence according to whether it is cloning or sequencing vector, *E.coli* contamination and sequence of an unacceptably poor quality. Clipped good quality sequences are then passed into the ‘Phrap2Gap’ program (Mott and Dear, unpublished), a modification of ‘Phred’ (a base-calling program) and ‘Phrap’ (a sequence assembly program; Gordon *et al*, 1998). ‘Phrap2Gap’ allows phrap-assembled reads to be transferred into the ‘GAP’ editing package. The ‘GAP’ sequence assembly program was developed as part of the Staden package (Bonfield *et al*, 1995; Staden, 1980; Staden *et al*, 2000); over the years versions have been updated from ‘xGAP’ to ‘GAP’ to ‘GAP4’ to ‘GAP4.new’. Clones assembled as part of this project were largely assembled using ‘GAP4.new’ packages.

2.7.5 Contiguation or ‘finishing’ of a clone

Generally, a clone is not a contiguous piece of DNA sequence upon transfer into a

'GAP' package. The clones, bA544A12 and bA18B16, were not contiguous and a number of steps were required in order to produce a 'finished' clone (defined as a contiguous piece of sequence with both cloning vector arms present). A 'finished' clone also required that all the sequence was 'double stranded', which refers to the idea that the entire clone should be covered by at least two individual reads. Assembling a clone, therefore, required the use of a number of pieces of software, resequencing certain subclones and generating specific segments of DNA using the PCR reaction with the addition of reaction additives (section 2.7.5.1).

After a clone was contiguous and double stranded, the virtual restriction digest of the clone was checked against fragments generated by 3 actual restriction digests. This involved generating the real digests (described in section 2.4) and generating the virtual digests. Virtual digests were generated by the program 'Confirm' (Production Software Group, Sanger Institute, unpublished) which also has a graphical display showing the real and virtual digests alongside each other.

2.7.5.1 'Finishing' PCR reaction

The additives A, E and F (Invitrogen) are used to sequence 'difficult' regions. Additive A is the 'universal additive' and is designed to generally aid the sequencing reactions on problematic areas. Additive E is used to sequence regions with high GA composition and additive F for high AT composition. The dGTP BigDye™ terminator mix (Applied Biosciences) is used for regions of high GC content.

1. 2 μl (40 nM) of forward primer and 2 μl (40 nM) reverse primer were dispensed into a 96 well plate (Costar), spun briefly and dried down in a 90°C

oven for 10 minutes.

2. To 3 µl DNA template, 4µl BigDye™ Terminator mix (Applied Biosystems) or dGTP BigDye™ Terminator Ready Reaction mix (Applied Biosystems), 2 µl additive A, E or F (Invitrogen) and 4 µl ddH_20 was added.
3. Samples were placed on the PTD-225 Peltier Thermocycler (MJ Research) with the following program: (i) 95°C for 15 seconds, (ii) 45°C for 5 seconds, (iii) 60°C for 2 minutes, (iv) steps (i)-(iii) repeated 25 times, (v) 4°C until program stopped.
4. Sequencing protocols were performed as described in section 2.7.2, using the appropriate primer(s).

Expression profile analysis

2.8 Design of paralogue specific primers

Primers were designed manually and using the Primer3 program (Rozen and Skaletsky, 2000) for each paralogue[1] to ensure that the primers and product were 'paralogue specific'[2]. Primer sequences, 18-25 bp in length with an average GC-content of 40-60% and melting temperature of 55°C-65°C, were designed, in most cases, in the 3' UTR of the paralogue mRNA to generate a PCR product between 250-500 bp. The primer sequences used are given in Appendix 3 and 4.

Sequences were chosen:

(i) to avoid areas of simple sequence showing non-representative use of the bases and obvious repetitive sequence i.e. runs of single nucleotides (e.g. TTTT) or double nucleotides (e.g. CGCGCG) motifs.

(ii) to avoid complementarity between primer pairs as this would result in primers annealing to each other and forming primer dimers.

(iii) to exclude palindromes which will form inhibitory secondary structure (e.g. GACGTC)

Each primer was also designed with the universal 5' adaptor sequence '5'-TGACCATG-3'' necessary for attaching the paralogue specific amplicon to the

[1] Paralogues (or paralogous genes) are genes found within the same species which have arisen by duplication of a common ancestral gene.

[2] 'Paralogue specific' indicates that the PCR primers and product have been designed to be specific to a particular paralogue and not to cross-hybridise with other members of the same paralogous gene family which might share high sequence homology.

surface of the microarray (as per section 2.13.2).

The specificity of each primer and product was determined by BLAST searching the sequence against the ENSEMBL human genome build (UCSC (ENSEMBL 1.1.0). Each PCR product was also verified by sequencing (section 2.7 using appropriate primer).

2.9 PCR amplification of paralogue specific PCR products

The primers used to amplify the paralogue specific PCR products for the Southern, Northern and dot-blot experiments are summarised in Appendix 3 and the primers used in the RT-PCR and microarray experiments are summarised in Appendix 4.

1. To 5 µl human genomic DNA (1µg/µl), 2 µl 10x PCR Buffer, 10mM dNTPs, 0.5 µl of primer 1 (200ng/µl), 0.5 µl primer 2 (200ng/µl), 0.125 µl Taq Polymerase and 10.875 µl of ddH_20 was added.
2. The paralogue-specific PCR products were amplified using a PTD-225 Peltier Thermocycler (MJ Research) with the following program: (i) 95°C for 5 minutes, (ii) 95°C for 1 minute (iii) [annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.
3. 5 µl of PCR product with 10 µl loading buffer were separated on a 2.5% agarose gel made up with 1x TBE and visualised with ethidium bromide.

2.10 Total RNA extraction from mammalian cell-lines

The 5 cell-lines growth medium (with serum and antibiotics): RPMI 1640 Medium

(GIBCO) supplemented with 10% fetal calf serum (FCS, GIBCO) and 5 ml penicillin/streptomycin (10,000U/ml; GIBCO BRL). Stored at 4°C.

1. The cell-line liquid nitrogen stocks were first thawed then washed by adding 25 ml of the tissue-culture medium and gently mixed.
2. The cell pellet was collected by centrifugation at 1500 rpm for 5 minutes and the supernatant discarded.
3. The cell pellet was resuspended in 15 ml of growth medium and grown in suspension at 37°C in 5% $C0_2$ / 95% air in 75 cm^2 filter capped flasks.
4. A cell culture with 70-80% confluence (~10^7 cells) was taken and centrifuged to pellet the cells at 1500 rpm
5. The medium was removed with a pasteur pipette and the cells washed twice with 50 ml PBS and the cell pellet collected by centrifugation at 1000rpm for 5 minutes.
6. 1ml TRIZOL reagent (GIBCO BRL) was added to each pellet and mixed well by pipetting until the pellet was completely resuspended (TRIZOL is a clear red liquid which will become cloudy once pellet is resuspended). An additional 1 ml TRIZOL was added if pellet did not resuspend completely.
7. The samples were dispensed into 1 ml aliquots in 2 ml round-bottom tubes and incubated at 60°C (heating block) for 10mins to fully resuspend the pellet.
8. 200µl chloroform was added to each 1 ml aliquot and mixed vigorously by shaking for ~15 seconds then incubated at room temperature for 2-3mins.
9. The samples were centrifuged at 14,000rpm for 15mins at 4°C.
10. The aqueous upper phase (clear, colourless) was dispensed into a new 2ml tube without disturbing the other layers and the remaining layers were discarded.

11. 0.5ml isopropanol was added to the aqueous layer and mixed by inversion then incubated at room temperature for 10mins.
12. The RNA pellet was collected by centrifugation at 14,000rpm for 15mins at 4°C (the RNA was visible as a white pellet at the bottom of the tube).
13. The supernatant was removed and discarded and the pellet washed once with 1 ml 75% ethanol. Vortexed to mix and centrifuged at 7,500rpm for 5mins at room temperature
14. The supernatant was removed and the pellet was centrifuged at 7,500 rpm for 2 minutes and the remaining supernatant carefully removed.
15. The pellet was air dried for ~30mins.
16. The pellet was re-suspended in 100µl DEPC and incubated at 60°C (heating block) to ensure the pellet was completely resuspended.
17. The total RNA was quantitated using a spectrophotometer and qualitated by assessing 2µg RNA by electrophoresis on a 1% agarose gel made with DEPC/1 x TBE (not exceeding 80mA as RNA will smear).
18. 3x volume 75% ethanol was added to the RNA sample. Stored at -70°C.

2.11 DNase treatment of RNA

Prior to use the total cell-line RNA was treated with DNase to remove any DNA contamination using the Ambion DNA-free™ kit.

1. 1 µg of RNA was incubated with 0.1 volumes 10 x DNase I Buffer (provided with kit) and 1 µl DNase I (2 units) at 37°C for 30 minutes.
2. 5 µl of DNase Inactivation Reagent (provided with kit) was added to the reaction mix and incubated at room temperature for 2 minutes.

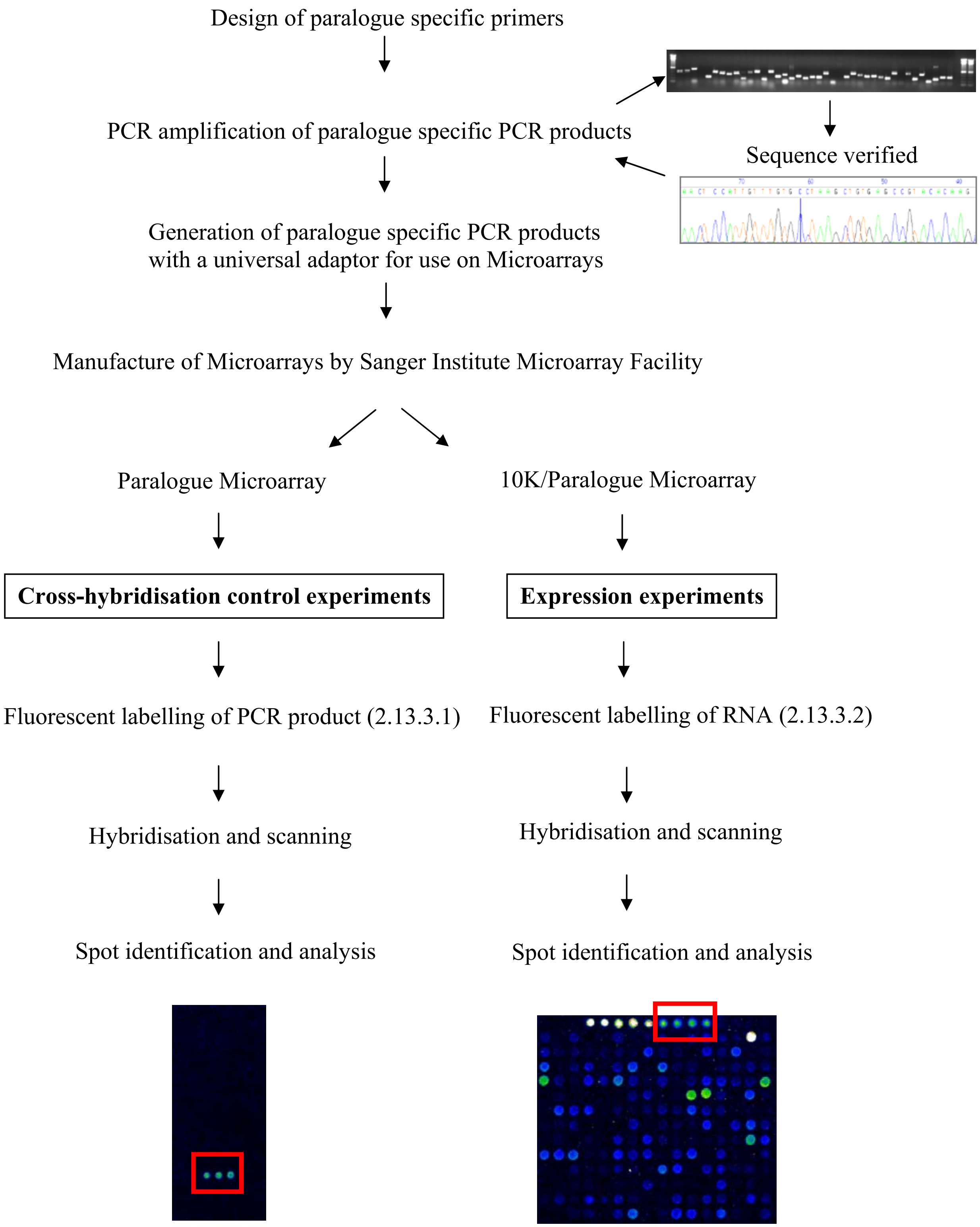
3. The DNase Inactivation Reagent pellet was collected by centrifugation at 13,000rpm for 1 minute and the supernatant containing the DNA free RNA removed into a fresh tube.

2.12 First strand cDNA synthesis and amplification of target cDNA using paralogue specific primers

The cDNA was synthesised using Superscript™ First-Strand Synthesis System for RT-PCR (Invitrogen). All reagents were provided with the kit.

1. To 1 µg DNA free total RNA 10mM dNTP mix, 1 µl Oligo$(dT)_{12\text{-}18}$ (0.5 µg/µl) was added and made-up to 10 µl with DEPC-treated water.
2. The reaction mix was incubated at 65°C for 5 minutes, then snap chilled on ice for 1 minute.
3. To the reaction mix, 2 µl 10x RT buffer, 25 mM $MgCl_2$, 0.1 M DTT and 1 µl RNaseOUT™ Recombinant RNase Inhibitor was added then incubated at 42°C for 2 minutes.
4. 1 µl of SuperScript™ II RT (50 units) was added to the reaction, mixed and incubated at 42°C for 50 minutes.
5. The reaction was terminated by incubating at 70°C for 15 minutes then chilled on ice.
6. The reaction was collected by centrifugation and 1 µl RNase H was added. The reaction was incubated at 37°C for 20 minutes.
7. Amplification of target cDNA was carried out according to section 2.9 substituting 5 µl genomic DNA with 2µl first-strand synthesised cDNA and increasing ddH_20 from 10.375µl to 13.375µl.

2.13 Overview of microarray experiments



Design of paralogue specific primers

PCR amplification of paralogue specific PCR products

Sequence verified

Generation of paralogue specific PCR products with a universal adaptor for use on Microarrays

Manufacture of Microarrays by Sanger Institute Microarray Facility

Paralogue Microarray

10K/Paralogue Microarray

Cross-hybridisation control experiments

Expression experiments

Fluorescent labelling of PCR product (2.13.3.1)

Fluorescent labelling of RNA (2.13.3.2)

Hybridisation and scanning

Hybridisation and scanning

Spot identification and analysis

Spot identification and analysis

2.13.1 Description of microarrays used

Two different microarrays were produced by the Sanger Institute Microarray Facility;

(i) The ‘Paralogue Microarray’ has 40 paralogue-specific PCR products arrayed in triplicate. This microarray was used to ensure that the amplified PCR products do not cross-hybridise with any of the other paralogue-specific PCR products.

(ii) The ‘10K/Paralogue Microarray’ is a modification of the Sanger Institute human 10K microarray (Hver1.2.1). Further information can be found at http://www.sanger.ac.uk/Projects/Microarrays. The 10K array consists of 12 x 4 super-arrays corresponding to 48 sub-arrays each containing 224 DNA elements arranged as 14 rows and 16 columns (figure 2.1). There are currently 9932 DNA elements corresponding to human genes on the 10K microarray. The first row of each sub-array is generally reserved for duplicate sets of positive and negative controls consisting of. Cy3 positive control (a spot of Cy3), a negative control (empty spot) and 5 bacterial controls representing *Bacillus subtilis trp*, *lysA*, *thrB*, *dapB* and *pheB* genes. The duplicate set of controls has been substituted on the 10K/Paralogue Microarray by 2 of the 40 paralogue-specific PCR products arrayed in quadruplicate. The quadruplicate sets of the 40 paralogues are represented in 2 different locations on the microarray; once in super-array rows 1-6 and again in super-array rows 7-12. Therefore, each paralogue specific PCR product appears 8 times on the microarray. The 10K/Paralogue Microarrays were used to determine the expression profiles of the paralogues and the standard 10K DNA elements in

10 human RNAs.

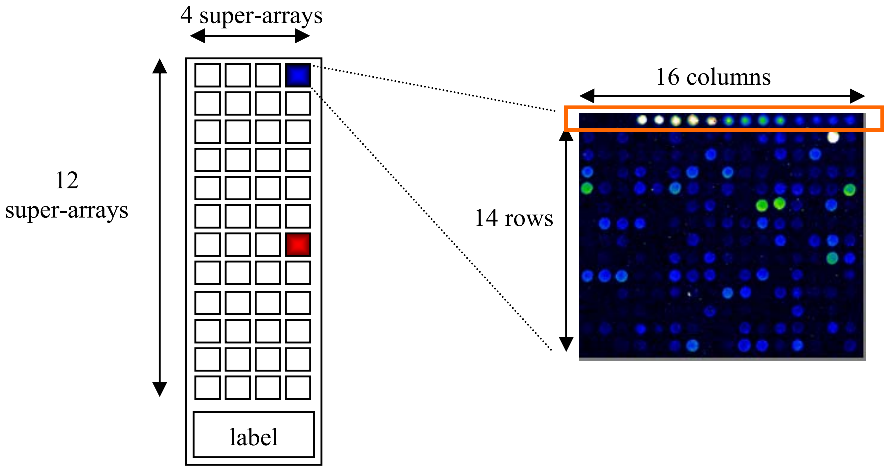


Figure 2.1 The 10K/Paralogue Microarray. The layout of the 48 sub-arrays in 12 x 4 super-arrays is shown and the sub-array coloured blue is expanded. The first row of the sub-array is boxed in orange. Columns 1 to 8 of row 1 contain the controls described in the text. The paralogue specific PCR products of one paralogue are arrayed in rows 9 to 12 (shown as 4 green spots) and a second paralogue in rows 13 to 16 (shown as 4 blue spots). The paralogue specific PCR products are represented in two different locations on the microarray and the super-array containing the same paralogue specific PCR products is coloured red.

2.13.2 Generation of paralogue specific PCR products with a Universal Adaptor for use on microarrays

The paralogue specific PCR products generated as described in section 2.9 were subjected to a second round of PCR in which a universal primer (5’ GCTGAACAGCTATGACCATG-3’) was used to attach an aminolinker to the 5’ of the PCR product. The aminolinker enables the attachment of the PCR product to the microarray surface.

1. The paralogue specific PCR products were amplified according to section 2.9. The bands corresponding to the required PCR products were excised from the gel and transferred into 1 ml $T_{0.1}E$ and placed at 4°C for 18 hours, then stored

at -20°C.

2. To 15µl of paralogue specific PCR product in $T_{0.1}E$, 6µl 10x PCR buffer, 3µl 10mM dNTPs, 1.5µl universal primer, 1.5µl paralogue specific reverse primer, 0.375µl Taq polymerase and 32.625µl ddH_20 was added.
3. The thermocycler program was as follows (i) 95°C for 5 minutes, (ii) 95°C for 1 minute (iii) [Annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.
4. 2µl PCR was analysed by electrophoresis on a 2.5% agarose/1 x TBE gel.
5. 15µl spotting buffer was added to each PCR product and this was arrayed onto the Microarrays.

2.13.3 Generation of fluorescently labelled DNA

2.13.3.1 Generation of fluorescently labelled paralogue-specific PCR products using the Cyanine 3-dCTP dye for hybridisation onto the 'Paralogue Microarray'

In order to ensure the specific PCR products do not cross-hybridise to the other paralogues the paralogue-specific PCR products generated in section 2.9 were labelled with a fluorescent dye and hybridised to the Paralogue Microarray.

1. To 5µl $T_{0.1}E$ DNA stocks of the paralogue specific PCR products 2µl 10x PCR buffer, 1µl 10 mM dA,T,GTP/5 mM dCTP mix, 0.5µl primer 1, 0.5µl primer 2, 0.125µl Taq polymerase, 2µl dCTP-Cy3 and 8.875µl ddH_20.
2. The thermocycler program was as follows (i) 95°C for 5 minutes, (ii) 95°C for

1 minute (iii) [Annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.

3. Excess nucleotides were removed from the PCR reaction using QIAquick Nucleotide Removal Kit (Qiagen) according to the manufacturers' instructions and eluted with 40µl ddH$_2$0.
4. 5µl of product with 5µl loading buffer were analysed on a 1% agarose 0.5 x TBE gel.
5. Depending on how successful the labelling reactions was, between 3-10µl of fluorescently labelled paralogue specific PCR product was denatured at 100°C for 5 minutes then snap chilled on ice and mixed with 38µl hybridisation buffer.

2.13.3.2 Generation of fluorescently labelled single-stranded cDNA target using direct incorporation of Cyanine dyes for hybridisation onto the '10K/Paralogue Microarray'

The Bacterial mRNA "cocktail" was provided by Sanger Institute Microarray Facility.

1. 1µl of bacterial "cocktail" (1 x stock in 75% ethanol) was added to 40µg of total RNA (in 75% ethanol) and precipitated by adding 1/40th volume of 3M sodium acetate at -70°C for 30 minutes.
2. The RNA pellet was collected by centrifugation at 13,000rpm and washed briefly in 100µl 70% ethanol and air-dried for 30 minutes.
3. The RNA pellet was resuspended in 12.9µl DEPC and 2.5µl anchored oligo-dT (2µg/µl final concentration; mixture of T$_{17}$A, T$_{17}$G and T$_{17}$C primers).
4. The RNA/oligo mixture was heated to 70°C for 10 minutes and then snap

chilled on ice.

5. To 15.4µl RNA/oligo mixture, 6µl 5x first strand buffer (Invitrogen), 3µl 0.1M DTT (Invitrogen), 0.6µl 10 mM dA,T,GTP/5 mM dCTP mix dNTPs, 3µl dCTP-Cy3 or dCTP-Cy5 and 2µl Superscript II (Invitrogen) was added.
6. The reaction was incubated at 42°C for 2 hours.
7. 1.5µl 1M NaOH was added to the reaction and incubated at 70°C for 20 minutes to hydrolyse the RNA.
8. 1.5µl 1M HCl was added to neutralise the reaction.
9. The nucleotides and short oligomers were removed using the Autoseq G-50 columns (Amersham Biosciences) according to the manufacturers' instructions resulting in ~33µl of labelled cDNA sample.
10. 33µl of test cDNA sample was combined with 33µl of control cDNA and 4µl polyA DNA (Sigma), 8µl C_0t1 DNA (Gibco BRL) and precipitated with 7.8µl 3M sodium acetate pH5.2 and 260µl 100% ethanol at -70°C for 25 minutes.
11. The pellet was collected by centrifugation at 13,000 rpm and washed briefly in 70% ethanol. All traces of ethanol were carefully removed and the pellet air-dried.
12. The pellet was resuspended in 40µl microarray hybridisation buffer and 8µl ddH_20.

2.13.4 Hybridisation, washing and scanning of microarrays

1. 46µl of hybridisation mixture was spotted onto the microarray cover slip (25 x 60 mm) and the microarray was inverted and lowered onto it.
2. The microarray was placed in a humid chamber (2 cm x 7 cm 3MM paper

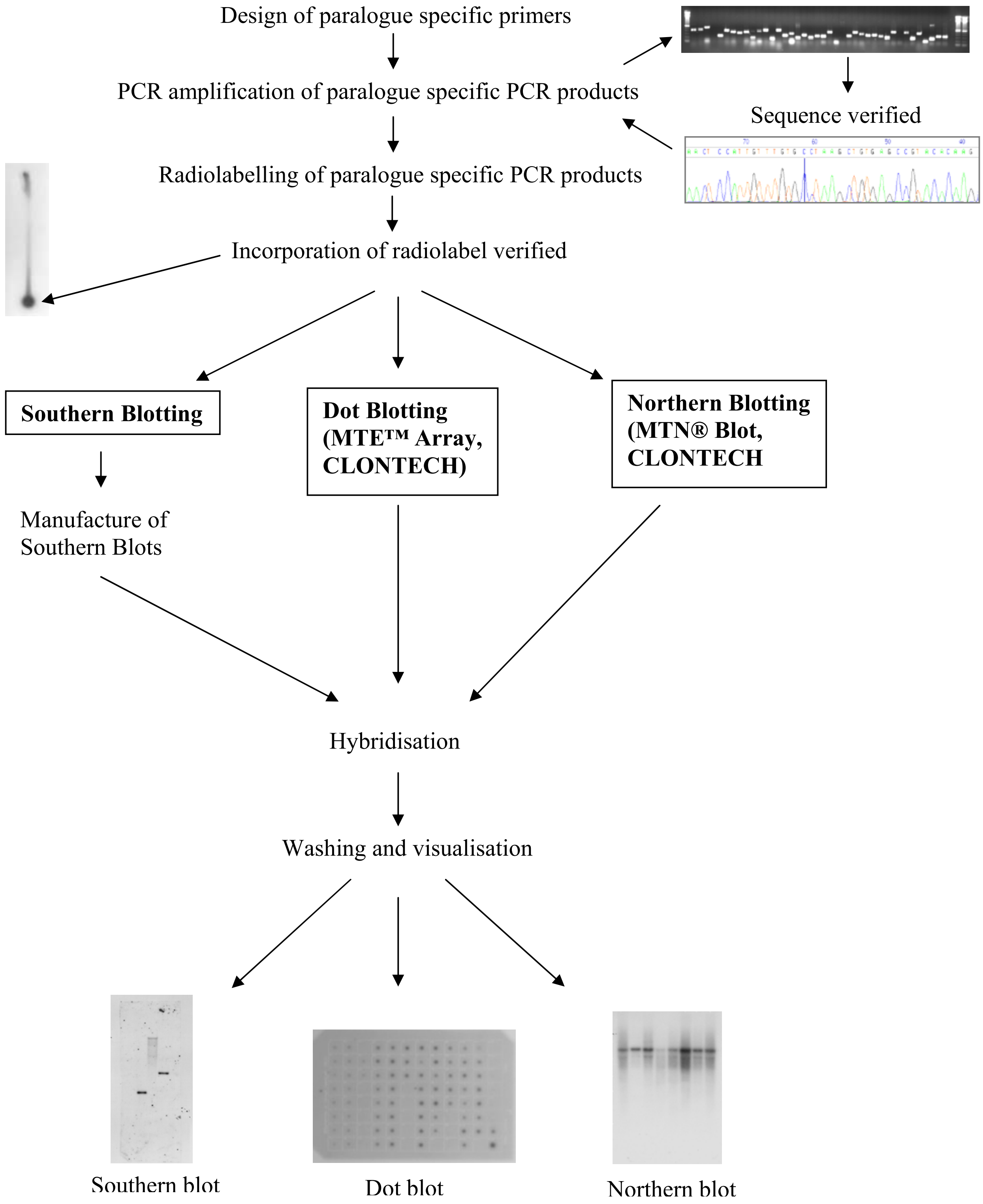
moistened with 2 ml 40% formamide, 2 x SSC in a Petri dish) and incubated for 12-24 hours at 47°C.

3. The cover slip was carefully removed from the microarray by rinsing in microarray wash solution 1 for 10-15 seconds.
4. The microarray was first washed in microarray wash solution 1 for 5 minutes at room temperature with gentle shaking. Followed by 2 washes in microarray wash solution 2 for 30 minutes at room temperature with vigorous shaking and, finally, in microarray wash solution 3 for 5 minutes with vigorous shaking at room temperature.
5. The microarray was dried by centrifugation at 1000 rpm for 1-2 minutes.
6. Using a laser-based scanner (GSI Lumonics ScanArray® 5000) the microarray was scanned at the two wavelengths compatible with efficient excitation for Cy3 and Cy5 (550nm and 650nm respectively) at 10 µm scanning resolution.

2.13.5 Analysis of microarrays

GSI Lumonics Quantarray® microarray analysis application software was used to determine the fluorescence intensity of spots in microarray images produced by ScanArray®. A three stage protocol was observed: (i) spot finding, (ii) spot quantitation; (iii) data export and visualisation. Once the spots have been identified and quantitated the standard deviation between the spot intensity and background intensity was calculated. In most cases, if a spot was present the standard deviation was greater than 2. To verify these results each spot was also assessed by-eye for each experiment using Quantarray®. The microarray data was clustered using the program EPCLUST at EMBL-EBI as described in section 2.18.

2.14 Overview of blot expression analysis



Design of paralogue specific primers

PCR amplification of paralogue specific PCR products

Sequence verified

Radiolabelling of paralogue specific PCR products

Incorporation of radiolabel verified

Southern Blotting

Dot Blotting (MTE™ Array, CLONTECH)

Northern Blotting (MTN® Blot, CLONTECH

Manufacture of Southern Blots

Hybridisation

Washing and visualisation

Southern blot

Dot blot

Northern blot

2.14.1 Radioactive labelling of DNA

2.14.1.1 Radioactive labelling of paralogue-specific PCR products

1. In a 0.5µl microcentrifuge tube, 2µl 10x PCR buffer, 1µl 10mM dNTPs mix, 0.5µl primer 1, 0.5µl primer 2, 0.125µl Taq polymerase, 4µl [α-^{32} P]-dCTP and 6.875µl ddH_20 was added to 5µl of the $T_{0.1}E$ DNA stocks of the paralogue specific PCR products generated as described in section 2.9.
2. The reaction was overlaid with mineral oil to prevent evaporation and subjected to PCR in a DNA thermal cycler (Perkin Elmer, USA). PCR cycling conditions were as follows (i) 95°C for 5 minutes, (ii) 95°C for 1 minute (iii) [Annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.
3. Excess nucleotides were removed using QIAquick Nucleotide Removal Kit (QIAGEN) according to manufacturers' instructions and the labelled PCR product was eluted in 50µl ddH_20.

2.14.1.2 Radioactive labelling of DNA using MegaPrime™ DNA labelling system (Amersham)

1. To 25 ng DNA template, 5 µl of primers was added and the final volume made up to 50 µl with ddH_20.
2. The reaction mix was denatured at 95 °C for 5 minutes.
3. The reaction was collected by centrifugation at 13,000 rpm.
4. 10 µl labelling buffer, 5 µl [α-^{32} P]-dCTP and 2 µl enzyme was added to the reaction mix. Mixed then centrifuged.

5. The reaction was incubated at 37°C for 1 hour and the reaction stopped by the addition of 5 µl 0.2 M EDTA.

2.14.2 Probe verification

2.14.2.1 Assessment of radiolabel incorporation using thin-layer chromatography

1. 1µl of PCR product was spotted onto a 5 x 10 cm Polygram CEL 300 PEI/UV thin-layer chromatography sheet (Macherey-Nagel, GmbH & Co) approximately 1.5 cm from the bottom edge.
2. This was placed in a beaker containing 0.75 M KH_2PO_4 (pH3.5) 1cm in depth and left for 30 minutes.
3. The chromatogram was subjected to autoradiography for 30 minutes.
4. Incorporated isotope remains at the spotting position, whereas unincorporated migrates with the buffer front.

2.14.2.2 Measurement of radioactively labelled PCR product concentration

The optimal concentration of radioactively labelled PCR product (or probe) is 1-2 x 10^7 cpm/ml. This was calculated using either a mini Geiger counter or a Scintillation counter (Easicount 4000, Scotlab, UK).

2.14.3 Manufacture of Southern Blots

2.14.3.1 Restriction digest of human genomic DNA

1. To 10µg human genomic DNA, 1mM Spermidine, 1mM DTT, 1.5µl restriction enzyme (either *Pst*I, *EcoR*I and *Hind*III), 5µl appropriate NEBuffer were added and the total reaction volume made up to 50µl with $T_{0.1}E$.
2. The reaction was incubated at 37°C for several hours (time optimised for each enzyme). After 1 hour another 1µl enzyme was added.
3. 3µl of digest was analysed on a 0.8% agarose 0.1 x TAE gel with a 100bp ladder. If digestion was not occurring after 24 hours more enzyme was added and the reaction incubated at 37°C until genomic DNA was completely digested and a further 3µl analysed by electrophoresis.
4. Once the genomic DNA had completely digested all 3 digests were loaded on a 0.8% agarose, 1 x TAE gel (2.4 g agarose, 300 ml 1 x TAE, 10 µl Ethidium Bromide) with a lambda *Hind* III ladder (100ng) and run at 50v for ~16 hours.

2.14.3.2 Transfer of digested genomic DNA onto filter

1. Excess gel was cut away and the gel was denatured in 1 x denaturation solution for 30 minutes with gentle shaking.
2. The gel was rinsed twice in ddH_20 then washed twice in 1 x neutralisation solution for 30 minutes each with gentle shaking.
3. Gels were blotted for 24 hours in 10 x SSC onto hybridisation transfer membrane (Hybond™-N ; Amersham) with frequent changing of towels.
4. The membranes were rinsed in 2 x SSC, dries on Whatman paper and the

DNA cross-linked on a UV transilluminator (320nm) for 2.5 minutes.

2.14.4 Hybridisation of radiolabelled PCR product to blots

1. The blots were prehybridised in ExpressHyb™ Hybridisation Solution (Clontech) containing 1.5 mg sheared salmon testes DNA (Stratagene) at 65°C for 1-4 hours.
2. In the case of the Human Multiple Tissue Northern (MTN®) blots (Clontech) and the Southern blots, the radiolabelled DNA were denatured at 95-100°C for 10 minutes then snap chilled on ice for 10 minutes before being added to appropriate volume fresh ExpressHyb™ solution.
3. In the case of the Human Multiple Tissue Expression (MTE™) Arrays (Clontech) 30µg of C_ot-1 DNA, 150µg of sheared salmon testes DNA (Stratagene) and 50µl 20 x SSC were added. The reaction volume was made up to 200µl with ddH_20 then heated to 95-100°C for 5 minutes then incubated at 68°C for 30 minutes.
4. The pre-hybridising solution was discarded and replaced with fresh ExpressHyb™ solution containing 1.5 mg sheared salmon testes DNA and the denatured radiolabelled PCR product. All the blots were hybridised at 65°C for 16-18 hours.

2.14.5 Washing

1. The hybridisation solution was discarded and replaced with the appropriate wash solution I. The blots were rinsed 5 times with wash solution I before being washed in fresh wash solution I for 30-40 minutes (the wash solution

was replaced several times) at room temperature for MTN and Southern blots and 65°C for MTE Arrays.

2. Wash solution I was discarded and the MTN and Southern blots washed at room temperature in wash solution II and MTE array washed at 65°C until the background signal was significantly reduced and activity detected with a Geiger counter more specific (~5 cpm).
3. Excess liquid was removed and from the filters by laying them briefly onto Whatman 3MM paper. The filters were then subjected to autoradiography using pre-flashed film and intensifier screens at -70°C for 24 hours, 3 days in all cases, and longer if necessary.

2.15 Computational analysis

A multitude of bioinformatics programs were used in this thesis in order to identify and characterise genes; both in the annotation of genomic clones and the identification and characterisation of MHC paralogues. The individual tools are discussed in section 2.15.1 and the methods in which they were used for a particular analysis are discussed in later sections.

2.15.1 General programs used in this thesis

'BLAST' is an acronym for the basic alignment search tool (Altschul *et al*, 1990). The program has become widely used in DNA and protein database searches. It is based on measuring local similarity between sequences, calculated by the maximal segment pair (MSP) score. There are several types of BLAST searches available for both nucleotide and protein sequences. In general, 'tblastn' was used to search the protein sequence against the selected nucleotide database translated in all six reading frames. In addition, PSI-BLAST (Position Specific Iterated BLAST) was used in the identification of paralogues. This uses an iterative search in which sequences found in one round of searching are used to build a score model for the next round of searching thus identifying paralogues sharing weak sequence homology. The databases searched using BLAST were either stored at EMBL-EBI or the NCBI.

'NIX' is a tool at the HGMP used to view the results of running many DNA analysis programs on a DNA sequence. In the initial step the sequence is masked for repeats using 'RepeatMasker' (Smit and Green, unpublished). This program screens against a library of interspersed repeats and low complexity DNA sequence called 'Repbase'

(Jurka, 2000). BLAST searches are started using the masked sequence against a number of databases, including Swissprot, TrEMBL, EMBL, EST, HTG, Unigene, Ecoli and Vector. The DNA sequence is also run through a number of gene finding programs, including ‘Grail’, ‘Genefinder’, ‘Hexon’ and ‘Fgene’. The results from so many different programs are presented in a graphical interface and the features in the DNA sequence identified. By viewing all the results side-by-side it makes it easier to see when many programs have a consensus about a feature.

‘Electronic PCR’ (e-PCR) is a tool used to identify molecular markers, such as STSs, in a query sequence. In order to determine the true locations of known genes on chromosome 9, e-PCR was performed using the cDNA sequences of the genes as the query sequences. The STSs matching the cDNA sequence of the gene were identified and used to determine which genomic clone the genes were located within the chromosome 9 database, ‘9ace’.

The ‘ENSEMBL’ genome browser was used extensively throughout this project. The ‘ENSEMBL’ genome browser is a joint project between EMBL-EBI and the Sanger Institute and provides a bioinformatics framework to organise biology around the sequences of large genomes (Hubbard *et al*, 2002). ENSEMBL provides a fully annotated human genome incorporating the data from existing biological databases and as *ab initio* gene predictions. Other genome browsers were also used to view the genomic sequence and annotation; the UCSC genome browser based at UC Santa Cruz (Kent and Haussler, 2001; Kent *et al*, unpublished) and the NCBI Map Viewer. All three genome browsers are now built using the same reference sequence constructed directly from the physical maps representing the minimum clone tiling path being finished by the genome centres. All analyses for this thesis were performed using NCBI31 genome data freeze (November 2002).

EMBOSS (Rice *et al*, 2000) is the 'European Molecular Biology Open Software Suite' which provides a comprehensive suite of sequence analysis programs (~100). Programs such as 'water' and 'needle' were used to generate local and global sequence alignments, respectively. Several other programs were used to manipulate both DNA and protein sequences.

2.16 Identification of extended MHC paralogous genes in the human genome

Chapter 4 summarises the results of a search to identify the extended MHC genes in the human genome. Several methods and programs have been developed to facilitate the identification of the paralogues but only the final method used to generate the results presented in chapter 4 is described here. The paralogues were identified with increasing levels of confidence using a number of criteria; the method used is described in sections 2.16.1 and 2.16.2.

2.16.1 Identification of extended MHC paralogues based on protein sequence homology

The extended MHC protein sequences were BLAST searched against the ENSEMBL human genome build NCBI31 using the 'tblastn' executable. The BLAST program used in this analysis was Washington University BLAST version 2.0 (WU-BLAST2.0) which is capable of detecting relationships between proteins with low sequence identities (Brenner *et al*, 1998). The BLAST search parameters have been optimised to identify the paralogues based on sequence homology and to eliminate false-positives and reduce background noise as much as possible without losing the sensitivity of the analysis. The 2 critical parameters optimised were; (i) the substitution matrix and (ii) the Expected (E) value.

i. The substitution matrix is a key element in evaluating the quality of an alignment and assigns a score for aligning any possible pair of residues. The BLOSUM62 (Henikoff and Henikoff, 1992) matrix was used in this analysis as it is one of the best for detecting weak sequence similarities of query length greater than 85 amino acids/nucleotides.
ii. The Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size (Karlin and Altschul, 1990). Essentially, the E value describes the random background noise that exists for matches between sequences and the E value is used as a convenient way to create a significance threshold for reporting results. The E-value used in this analysis was 10. Using a set of protein sequences I found that by increasing the E value from 10 a larger list with more low-scoring hits was reported and no new paralogues were identified. By decreasing the E-value the number of hits was reduced and the low-scoring hits were practically eliminated but known paralogues were also not identified.

The resulting BLAST hits were then filtered according to the P-value and results with a P-value $\geq E^{-5}$ removed from the analysis. The P value is the probability of finding such a hit by chance. In short, small P values are considered to be good and very unlikely to be random, therefore meaningful, but P values become unreliable above

10^{-5} (Lesk, 2002). This filtering value was selected in order to maintain both sensitivity and specificity of the experiment.

2.16.2 Identification of extended MHC paralogues with increasing levels of confidence

The initial BLAST search identified 1000's of BLAST hits. Using knowledge of the protein sequence and the gene structure these BLAST hits were filtered to identify the extended MHC paralogues with the highest level of confidence.

2.16.2.1 Filter 1: Domain-masking

The protein domains were identified by searching the protein sequence against the 'PFAM' database of protein domain families using the perl script 'pfam_scan.pl' (written by the PFAM Software Group and kindly provided by K.Howe). The domains were masked using another perl script 'x_out_domains.pl' (written by K.Howe). The domain-masked protein sequences were then BLAST searched against the human genome as described in section 2.16.1 and the results sorted according to the P value. By masking the protein domains a large number of BLAST hits were identified that were just to a particular protein domain. More significantly, it also identified the extended MHC paralogues which still share good sequence homology outside the domains.

2.16.2.2 Filter 2: FINEX

Putative paralogues were initially identified by sequence homology using similarity searching to find relationships. However, genomic sequence data provides gene architecture information not used by conventional search methods. In particular, intron positions and phases are expected to be relatively conserved features, because mis-splicing and reading frame shifts should be selected against. 'FINEX' (Fingerprinting of INtron EXon boundaries) is an alignment technique which exploits the gene structure information provided by a genomic sequence (Brown *et al*, 1995). A single exon fingerprint can be compared rapidly against all the entries in a library of fingerprints which is generated using the CDS (coding sequence information) features in the annotated EMBL entry (EMBL release 73). The phases of the exon fingerprints are classified according to their position relative to the reading frame of the gene: introns lying between two codons (phase 0); introns interrupting a codon between the first and second base (phase 1); and, introns interrupting a codon between the second and third base (phase 2). These intron positions and phases are expected to be relatively conserved features, because mis-splicing and reading frame shifts should be selected against.

The FINEX database relies on coding sequence (CDS feature) information available in annotated EMBL entries for genomic clones. Only a small percentage of genomic clones are annotated therefore the FINEX database does not contain the fingerprint for every gene in the genome. As the MHC region is one of the best characterised regions and the majority of clones covering the region are annotated all the MHC gene fingerprints are present in the FINEX database compiled using EMBL release version 73. Therefore, the FINEX fingerprint was generated for all putative paralogues and used to search the FINEX fingerprint database using the optimised parameters: weight = 0.5, power = 4.0 and gap penalty = 0.5.

A number of scores were generated for each alignment to add statistical significance. The important scores to consider when determining a cut-off threshold are the D_{avg} score, which is the alignment length normalised score used to rank the alignments, the D_{mat} score, which is the global alignment score that measures the quality of the alignment and the z-score, which is the significance of the D_{mat} score for a given query/hit. The best alignment attainable is with self and, by definition, the dissimilarity scores D_{mat} and D_{avg} are zero and has the highest z-score (Brown *et al*, 1995). With this in mind the putative paralogues were used to search the FINEX fingerprint database and it was generally observed that the highest z-score obtained corresponded to the MHC gene it is paralogous to, or to another paralogue. Analysis of the paralogues identified ambiguous matches with a z-score less than 3.00 therefore a z-score greater than 3.00 was taken as significant. This is in agreement with significant z-score values for conventional sequence comparison (Dayhoff *et al*, 1978; Feng *et al*, 1984).

2.17 *In-silico* expression analysis

The ‘UNIGENE’ database at the NCBI automatically clusters the ‘GenBank’ sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information including the tissue types in which the gene has been expressed. This information, in the form of expressed sequence tags (EST) was downloaded into a text file <gene_name>_unigene.txt for each gene. In addition to ensure no ESTs had been omitted the gene mRNA or cDNA was used to BLAST search the database ‘dbEST’; a division of GenBank that contains sequence data and other information on ‘single-pass’ cDNA sequences.

The ‘BLAST’ results were then filtered to remove insignificant hits (generally <90% sequence identity) and the results saved in a text file <gene_name>_dbest.txt. The 2 text files were then parsed to produce a list of EST accession numbers using ‘parse.pl’ and the 2 lists were compared to identify ‘unique’ and ‘not-unique’ ESTs for a particular gene using the perl script ‘check_unique.pl’. The ‘not-unique’ ESTs were removed from the list and the non-redundant EST lists for each paralogue were compared against each other and ‘not-unique’ ESTs removed resulting in a paralogue specific list of ESTs.

The DNA sequence for each paralogue specific EST was extracted using ‘sequence_retrieval.pl’ and aligned against the gene cDNA using the program ‘b2b’ (written by R.Horton) and any false positive ESTs removed from the analysis. A database containing the full database entry for each paralogue specific EST list was produced using ‘db_extract.pl’ and the tissue information extracted using ‘tissue.pl’. As the annotation within EST database entries is not consistent several versions of ‘tissue.pl’ exist to extract the tissue information. Finally, the EST data was inputted into Excel and sorted by tissue and system. All perl scripts used in this section were written by K.Crum and modified by myself, unless stated otherwise.

2.18 Clustering methods

The data presented in this thesis was clustered using the unsupervised clustering methods, hierarchical clustering (clustering methods are reviewed by Brazma and Vilo, 2001) using using EPCLUST (Expression Profile Data CLUSTering and Analysis) at the EBI. Hierarchical clustering arranges the data in a tree-like structure (similar to phylogenetic trees), where genes with similar expression patterns occupy neighbouring 'leaves' of the tree. The algorithms used for hierarchical clustering are largely the same as used for distance-based phylogenetic reconstruction from sequence data, but restricted to those methods that are fast enough to deal with large numbers of nodes. Hierarchical clustering works by iteratively partitioning clusters starting with the complete set. After the joining of two clusters, the distances between all other clusters and a new joined cluster are recalculated. The complete linkage method used in this analysis uses the maximum distances between the members of two clusters to cluster the data.

Expression data was also clustered using the perl script ‘exprofile’ (written by R.Younger), or modifications of this script. The input file consists of a tab delimited table containing the expression data. The cells contain values 0 and 100 which correspond to whether the gene is expressed in the corresponding tissue or not, respectively. Alternatively, the cells may contain a value between 0 and 100, which is the percentage of genes expressed in a particular tissue. The output of the program is a postscript file containing the image corresponding to the input data, i.e. where a gene is expressed there is a black bar and when there is no expression it is white. The thickness of the black bar corresponds to the percentage of genes expressed in a particular tissue. This is described in more detail in chapter 6.

2.19 Phylogenetic analysis

The consensus phylogenetic trees presented in Chapter 5 were produced by merging the trees generated by 3 different software packages; PHYLIP, TREE-PUZZLE and MEGA2. The PHYLIP (Phylogeny Interface Program version 3.6; Felsenstein, 1989) package is a public domain package that provides a wide range of programs for constructing phylogenetic trees from molecular and other types of data. TREE-PUZZLE version 5.0 (Schmidt *et al*, 2002) is a computer program to reconstruct phylogenetic trees from molecular sequence data by maximum likelihood using the quartet-puzzling algorithm and MEGA2 (Molecular Evolutionary Genetics Analysis) software (Kumar *et al*, 1994; Kumar *et al*, 2001) is a software package for exploring and analysing aligned DNA or protein sequences from an evolutionary prospective and offers useful and easy-to-use methods of comparative sequence analysis.

2.19.1 Protein sequence alignments

Protein sequences were aligned using the 'ClustalW' program (Thompson *et al*, 1994). This is a progressive multiple sequence alignment method which, firstly, assigns individual weights to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones. Secondly, it varies amino acid substitution matrices at different alignment stages according to the divergence of the sequences to be aligned. Thirdly, residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. After a gap has been opened, locally reduced gap penalties are applied to positions around this gap. The alignments produces by ‘ClustalW’ were viewed in ‘belvu’ (Sonnhammer, unpublished) and edited using ‘Jalview’ (Clamp, unpublished).

2.19.2 Estimation of the gamma distribution

Evolutionary analysis of DNA and protein sequences is typically performed by either assuming that all evolutionary lineages evolve at the same rate or by avoiding any attempt to directly consider the fact that the rate of evolution changes over time. The default parameters for the 3 programs used assume that the rate of evolution is constant. However, there are several factors that affect the rate of molecular evolution (e.g., mutation, population size, selection) and therefore the rate of molecular evolution is extremely unlikely to be identical for different evolutionary lineages or individual amino acids or nucleotides. This was taken into account in this analysis and the rate of variation (or rate of heterogeneity) between sites was calculated using the gamma distribution. The shape of this distribution is determined by the value of a parameter known as the gamma distribution parameter alpha and was calculated using TREE-PUZZLE.

2.19.3 Bootstrapping and tree-puzzling steps

The aligned sequences were bootstrapped using the program SEQBOOT in PHYLIP and by selecting the bootstrapping option in MEGA2. Bootstrapping (Felsenstein, 1985) involves taking each site within a protein and rearranging sites to create a number of ‘pseudoalignments’. These ‘pseudoalignments’ are then used to recreate a number of trees which are compared to the original tree. Groupings obtained in the original tree are then given a percentage expressing how many times they are recreated in the ‘pseudoalignment’ trees.

The ‘puzzling-step’ parameter was selected in TREE-PUZZLE which is similar to bootstrapping and trees are composed into so-called intermediate trees. This step results in many intermediate trees (default 1000) and from these a majority rule consensus tree is built and the number of intermediate trees lending support for the consensus topology is displayed at each node. Bootstrap or puzzling-step values of over 50 % were considered to represent reliable groupings those below were considered to show little or no support. However, low values at branches are not considered worthless as every phylogenetic tree is the best tree obtainable using a specific method and sequences. Computer simulations have shown that the branching patterns of an inferred tree may be correct even if they are not supported by high bootstrap values (Nei and Kumar, 2000).

2.19.4 Phylogenetic analysis using distance methods

Trees were generated using the Neighbour-Joining method (Saitou and Nei, 1987). This method uses an algorithm to convert pairwise distances between sequences into a matrix, from which branching order and branch lengths are computed. The Jones, Taylor and Thornton, or JTT, (Jones *et al*, 1992) model of amino acid change was used. This model is very similar to another model, the PAM Dayhoff (Dayhoff *et al*, 1978) model, which provides a measure of probability calculating how likely the amino acid in one sequence is likely to change the amino acid in the other sequence. These probabilities were based on a subset of closely related proteins that were organised into a phylogenetic tree and the frequency of change from each amino acid to another was determined by adding up the changes at each evolutionary step. The JTT model is based on a recounting of the number of observed changes in amino acids of a much larger set of proteins therefore this model is to be preferred over the original Dayhoff PAM model. Using this model the Neighbour-Joining method constructed trees from the matrices of the multiple data sets from bootstrapping by the successive clustering of lineages and the setting of branch lengths as the lineages join.

2.19.4.1 PHYLIP

The output of the SEQBOOT program in PHYLIP was used as the input into the distance program PROTDIST. The program corrects distances for unequal rates of change at different amino acid positions using the coefficient of variation (CV) which was calculated using the gamma distribution alpha parameter from TREE-PUZZLE. The square of the CV is the value of the alpha parameter. The PROTDIST output was used as an input file to the program NEIGHBOR. The consensus tree is produced by using the output of the NEIGHBOR program as the input of the CONSENSE program.

2.19.4.2 MEGA2

MEGA2 is an easy to use software package and phylogenetic trees are generated quickly and in one simple step. First, the protein alignments were converted into MEGA2 format (.meg file) within the software package then the trees were generated using the neighbour-joining method (Saitou and Nei, 1987) under the JTT (Jones *et al*, 1992) model with 1000 bootstraps. Two additional parameters were selected; the pairwise deletion comparison option and the Gamma distance option. The former removes sites containing missing data or alignment gaps from the analysis as they arise. This is in contrast to the complete-deletion option which removes all such sites prior to analysis. Both options were initially used but no significant difference was observed. The Gamma distance was used to take care of the inequality of the substitution rates among sites and the gamma shape parameter, or alpha parameter, calculated using TREE-PUZZLE was used in this analysis.

2.19.5 Phylogenetic analysis using the maximum likelihood method

The maximum likelihood (ML) method allows the inference of evolutionary trees from nucleotide or amino acid sequences under a probabilistic model of nucleotide/amino acid evolution (Felsenstein, 1981). The ML method looks for all possible tree topologies between the sequences by initially constructing an unrooted tree using three sequences then the 4th is added to the tree and the 'best' tree topology for the four sequences chosen under likelihood criterion. This is repeated for the 5th, 6th, 7th etc sequences until the final tree is produced. The log likelihood value is calculated and the best tree is the one with the most positive log likelihood value.

2.19.5.1 PHYLIP

The PHYLIP program PROML implements the maximum likelihood method for protein amino acid sequences using the JTT model of changes between amino acids. The model assumes that each position and each lineage have evolved independently and the different rates of evolution were determined using the Gamma distribution. As previously the gamma distribution alpha parameter was calculated using TREE-PUZZLE and the Coefficient of Variation used as input into PROML. PROML is CPU intensive and, for this reason, data analysed was not bootstrapped.

2.19.5.2 TREE-PUZZLE

TREE-PUZZLE (TREE-PUZZLE version 5.0) constructs phylogenetic trees using maximum likelihood by implementing the fast tree search algorithm, quartet-puzzling. The protein alignment was used as input into the program and the 1000 'puzzling-step' option selected. Trees were generated using the JTT model and the rate of heterogeneity was set as Gamma distance. The 'outfile' generated contained information regarding the calculation of the gamma distribution alpha parameter and the consensus tree.

2.20 Useful web-sites

BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
Chromosome 6	http://www.sanger.ac.uk/HGP/Chr6/
Chromosome 9	http://www.sanger.ac.uk/HGP/Chr9/
ClustalW	http://www.ebi.ac.uk/clustalw/index.html
Electronic PCR	http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
EMBL-EBI	http://www.ebi.ac.uk
EMBOSS	http://www.hgmp.mrc.ac.uk/Software/EMBOSS/overview.html
ENSEMBL	http://www.ensembl.org/Homo_sapiens
EPCLUST	http://www.ebi.ac.uk/microarray/ExpressionProfiler/ep.html

FINEX	http://www.sanger.ac.uk/cgi-bin/finex/finex_search.pl
GeneMap99	http://www.ncbi.nlm.nih.gov/genemap99/
HGMP	http://www.hgmp.mrc.ac.uk
HUGO	http://www.gene.ucl.ac.uk/hugo/
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink
MEGA2	http://www.megasoftware.net
MIPS	http://mips.gsf.de
NCBI	http://www.ncbi.nlm.nih.gov
NIX	http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/
PFAM	http://www.sanger.ac.uk/Software/pfamservice.shtml
Primer3	http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
Repbase	http://www.geospiza.com/products/tools/repbase.htm
RepeatMasker	http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
Sanger Institute	http://www.sanger.ac.uk
TREE-PUZZLE	http://www.tree-puzzle.de
UniGene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene