Chapter 2

Materials and Methods

2.1. Materials

The majority of chemical reagents were bought from Sigma unless stated in the text; similarly restriction enzymes were largely bought from New England Biolabs unless stated elsewhere. A number of kits from various companies were used: these are specified in the text. PCR was generally performed using Amplitaq and the supplied PCR buffer from Perkin Elmer unless otherwise stated. All primers used in this thesis are listed in Appendix 1.

In the list of materials, where materials have an ambiguous name the numbers in brackets refer to the section for which the material is required.

2.1.1. Solutions

[2.2] Solution I (GTE/GET): 50 mM Glucose, 25 mM Tris, 1 mM EDTA (2.3 ml 20% glucose, 5.0 ml 0.1M EDTA, 1.3 ml 1M Tris (pH 7.4), 42 ml water)

[2.2] Solution II (NaOH/SDS): 0.2 M NaOH, 1% SDS (2.5 ml 4M NaOH, 2.5 ml 20% SDS, 45 ml water.)

[2.2] Solution III: 3.5 M KOAc (pH 5.5) (147.21g potassium acetate, 57.5 ml glacial acetic acid, water to 500 ml)

[2.2] Solution IV: TE (10:0.1) with RNase (10 μg/ml)

Sodium acetate/EDTA solution: 49.218g sodium acetate, 2 ml 0.1 M EDTA, water to 200 ml.

[2.3] Hybridisation solution: 50% formamide, 2 x SSC (pH 7.0), 10% dextran sulphate, 1% Tween 20

SSCTM: 4 x SSC (pH 7.0), 0.05% Tween 20, 5% low-fat dried milk

Buffered phenol: 1 ml phenol, 200 µl 1M Tris-hydrogen chloride (Shaken and placed on ice for 5 minutes, spun, top layer removed and discarded, 200 µl TE (10:0.1) added, mix shaken and spun. Kept on ice until required.)

[2.5.1] *EtOH/NaOAC mix:* 100 ml sodium acetate, 1600 ml ethanol 96%, 300 ml water.

RNase A: 200 mg RNAse A, 100 µl Tris (pH 7.4), 150 µl sodium chloride, water to 10 ml.

1 mM Tris-HCl (pH 8.5): 0.0606 g Tris, made up to 500 ml with water, and adjusted to pH 8.5 with hydrochloric acid.

[2.7] *Denaturing solution:* 25 ml of 10 M sodium hydroxide, 150 ml of 5 M sodium chloride, made up to 500 ml with water.

[2.7] 10 x Neutralization solution: 250 ml of 1 M Tris-chloride, 150 ml of 5 M sodium chloride, made up to 500 ml with water.

IPTG (0.1 M): 0.238 g in 10 ml of water. Sterilized by filtration, then stored at -20°C.

Xgal: 20 mg/ml dissolved in dimethyl sulfoxide (DMSO).

Trypsin-EDTA: 6.4 g sodium chloride, 0.16 g potassium chloride, 0.92 g sodium phosphate, 0.16 g potassium dihydrogen phosphate, 0.16 g sodium-EDTA, 0.5 g trypsin, made up to 1 litre with water. Stored at -20°C.

PBS: 10 g sodium chloride, 0.25 g potassium chloride, 1.44g sodium hydrogen phosphate (dibasic), 0.25 g potassium dihydrogen phosphate, made up to 1 litre with water and made to pH 7.4 with sodium hydroxide. Stored at 4°C.

Cresol Red Solution: 84.5 mg Cresol red sodium salts (Aldrich) in 100 ml TE (10:0.1). Stored at -20°C.

34.6% Sucrose: 121.1 g sucrose dissolved in 350 ml water. Stored at -20°C

10 mM dNTP: 1000 µl of each 100 mM dNTP (Amersham Pharmacia), 6 ml water. Stored at -20°C.

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

2 x SSC, 0.1% SDS: 100 ml 20 x SSC, 1 g SDS, made up to 1 litre with water.

0.2 x SSC, 0.1% SDS: 10 ml 20 x SSC, 1 g SDS, made up to 1 litre with water.

2.1.2. Media

Circlegrow: from QBiogene

*2 x TY:*15 mg/ml bacto-tryptone, 10 mg/ml bacto-yeast extract, 5 mg/ml NaCl (pH 7.4), water up to 1 litre.

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

SOC: SOB + 200 µl 20% glucose.

TYE plates: 8 g tryptone, 5 g yeast extract, 8 g sodium chloride, 12 g agar, water up to 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. (Final concentration of 50 µg/ml).

H-Top: 8 g Bacto Agar, 10 g Bacto Tryptone, 8 g sodium chloride, made up to 1 litre with water.

LB medium: 10 g Bacto-tryptone, 5 g Bacto-yeast extract, 10 g sodium chloride, made up to 1 litre with water, and pH 7.5 with sodium hydroxide.

LB/ampicillin/IPTG/X-Gal plates: 10 g Bacto-tryptone, 5 g Bacto-yeast extract, 10 g sodium chloride, 15 g Bacto-Agar, made up to 1 litre with water. 2 µl ampicillin (25 mg/ml) was added after the agar had cooled to 48°C. 1 ml of 0.1 mM IPTG and 2 ml Xgal (40 ug/ml) were also added to the cooled media, before plating out.

2.1.3. Loading dyes

Blue dextran formamide dye: 9.8 ml deionised formamide, 200 µl of 0.5 M EDTA, 0.01 g of blue dextran

[2.4] Loading dye: 5 mg bromophenol blue, 0.5 g Ficoll 400, 0.5 ml 10 x TBE, 4.5 ml water.

Ficoll dye: 0.5 g Ficoll 400, 100 µl 50 x TAE, 10 mg bromophenol blue, 4.9 ml water.

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

2.1.4. Buffers

TE (10:1): 2 ml Tris (pH 7.4), 2 ml 0.1 M EDTA, water to 200 ml.

TE (10:0.1): 2 ml Tris (pH 7.4), 200 µl 0.1 M EDTA, water to 200 ml.

10 x TBE buffer, pH 8.8: 162 g Tris base, 27.5 g boric acid, 9.2 g EDTA (disodium salt – Na_2EDTA), made up to 1 litre with water.

[2.4] Loading buffer: 10 µl 10 x TBE, 20 µl loading dye, 50 µl water.

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

50 x TAE buffer: 121 g Tris base, 12.2 g EDTA (disodium salt – Na_2EDTA), 28.55 ml acetic acid, made up to 500 ml with water.

10 x Rxn buffer: 4.5 ml 1M Tris-HCl (pH 8.8), 5 ml cresol red solution, 0.15 ml water, 0.35 ml 1 M magnesium chloride (BDH), 0.1454 g ammonium sulphate (GIBCO).

[2.12] Dilution buffer: 100 ml water, 50 ml TE (10:0.1), 0.8125 ml Cresol Red solution, 50 µl 4 M sodium hydroxide. Stored at -20°C.

2.1.5. Markers

[2.4] λ Hind III marker: 8 µl λ DNA- Hind III digest (NEB), 60 µl TBE buffer, 252 µl water. (Mixture was incubated at 65°C for 5 minutes before being rapidly chilled on ice.

80 µl of loading dye was then added.)

[2.4] pBR322 marker: 4 µl pBR322 DNA-BstNI digest (NEB), 60 µl 10 x TBE buffer, 256 µl water, 80 µl loading dye.

1 Kb ladder marker: 5 µl 1 Kb ladder mixed with 1 µl 50 x TAE, 10 µl Ficoll dye, and 34 µl water

[2.6.2] λBst11071 ladder marker: Prepare λBst11071 digest: 2 µl λ DNA (1 µg), 5 µl 10x buffer, 1 µl Bst11071 enzyme. Incubate at 37°C for 2 hours.

[2.6.2] Ladder marker: 2 µl 1 Kb ladder DNA (Gibco BRL, 1 µg/µl), 12 µl 50 x TAE, 50 µl λBst11071 digest, 100 µl Ficoll dye, 436 µl water.

2.1.6 Sequencing gel

Denaturing acrylamide gel (6%): 30 g urea was placed in a 250 ml beaker, together with 9 ml acrylamide/bisacrylamide solution, 4 ml 10x TBE and 37 ml water. The urea was dissolved by heating (60°C) and stirring, and the solution was made up to 60 ml with water. The solution was then placed in a dessicator for 4 minutes and just before pouring the gel 138 µl of 25% ammonium persulphate and 138 µl TEMED were added. The gel mix was then syringed between the glass gel plates whilst tapping the glass gently to get rid of the bubbles. The gel was left to set for at least 90 minutes prior to use.

2.1.7. Antibiotics

Ampicillin (stock solution of 25 mg/ml in water stored at -20°C): add to final concentration of 50 µg/ml

Chloramphenicol (34 mg/ml in 100% ethanol stored at -20°C): add to final concentration of 30 µg/ml

Kanamycin (25 mg/ml in water stored at -20°C): add to final concentration of 50 µg/ml

Tetracycline (12.5 mg/ml dissolved in 50% ethanol, stored in the dark at -20°C): add to final concentration of 15 μg/ml (used with media without magnesium salts).

500x Penicillin/Streptomycin (Boehringer Mannheim): Penicillin (50000 IU/ml), Streptomycin (50 mg/ml). Stored at -20°C.

2.1.8. Web addresses and other *in silico* resources

A number of the methods described require the use of various websites. In the text this is shown by the program name being in italics: websites used in the course of this thesis are listed in Appendix 3. A number of UNIX-based programs were also used in the course of this thesis: these are referenced in the text or listed in Appendix 4. A number of the programs listed in Appendix 4 were written by me in the Perl programming language.

2.2. Fluorescent Fingerprinting

With the advent of large-scale genomic sequencing, it was necessary to develop a method for producing large, deep contig maps allowing an optimal tiling path to be chosen. One method used to achieve this type of contig map is restriction digest fluorescent fingerprinting: this is based on using fluorescently tagged dideoxy ATPs to label the HindIII termini created in a double digest of the clone with HindIII and Sau3AI and using restriction patterns to assess the degree of overlap (Gregory *et al.*, 1997). The mouse MHC-linked OR contig was fingerprinted using this fluoresecent fingerprinting method.

2.2.1. Preparation of DNA for fluorescent fingerprinting (based on Birnboim and Doly, (1979))

1. 500 µl of 2 x TY (containing the appropriate antibiotic, kanamycin for PACs and chloramphenicol for BACS) was dispensed into four 96-well 1 ml Beckman boxes.
2. From the colonies that had been grown, clones were picked into these wells using sterilised toothpicks, and boxes were placed in a shaker at 300 rpm, 37°C for 12-18 hours.
3. 250 µl of each of the cultures were transferred to a round-bottomed 96-well Corning plate using a 50- to 250-multichannel pipette (Finnpipette) and cells were pelleted by centrifugation at 2500 rpm, 20°C for 4 minutes.
4. The supernatant was discarded from the cell pellets and pellets were resuspended in 25 µl of solution I and 25 µl of solution II. (Mixed by tapping the plates gently and then leaving the plates for 5 minutes.)
5. The supernatant was discarded once again and 25 µl of chilled solution III was added before leaving the plate for 5 minutes.
6. Well contents from the plates was transferred to filter plates (Millipore) taped to a round-bottomed Corning plate containing 100 µl isopropanol.
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 10 minutes.
9. The supernatant was discarded from the plate and the DNA pellet was briefly dried before being washed with 100 µl of 70% ethanol, centrifuging at 3200 rpm 20°C for 10 minutes.

10. Finally, the supernatant was discarded and the DNA pellet was dried before being resuspended in 5 µl of solution IV.

2.2.2. Generation of fluorescent marker

1. To make the fluorescent marker, 1.5 µg of lambda DNA (500 ng/µl) was placed in a 1.5 ml eppendorf tube with 7.5 U of BsaJI (2.5 U/µl), 16 U of TaqFS (8 U/µl), 2 µl of ROX ddC (5.08 µM), 5 µl of NEB2 buffer, and 35 µl of TE (10:0.1) (pH 7.4).
2. This tube was then incubated at 60°C for an hour before adding 50 µl of sodium acetate (0.3M) and 200 µl of 96% ethanol, mixing the solutions by vortexing briefly, and then leaving the tube (in the dark) at room temperature for 15 minutes.
3. The tube was then left at -20°C for a further 20 minutes before centrifugation at 14000 rpm for 20 minutes.
4. The supernatant was discarded, and the pellet was air-dried before 100 µl of 70% ethanol was added and the tube was again spun at 14000 rpm for 20 minutes.
5. After spinning, the supernatant was discarded and, after drying, the pellet was resuspended in 60 µl of TE (10:0.1) (pH 7.4) and 60 µl of blue dextran formamide dye.

2.2.3.The restriction digest reaction

1. A mix for the fluorescent labelling reaction was made up. The constitution of this was calculated by considering the amount of chemicals required per fluorescent labelling reaction, namely: (i) 2.8 U of HindIII (20 U/µl), (ii) 3 U of Sau3AI (30 U/µl), (iii) 3.7 U

of ThermoSequenase (32 U/μl), (iv) 0.14 μl of fluorescent ddA (10 μM; either HEX, TET or NED), (v) 0.8 μl of TE (10:0.1), (vi) 1 x NEB2 buffer.

2. Using a Hamilton repeat dispenser, 20 μl of this reaction mix was added to wells containing 5 μl of DNA suspended in TE (10:0.1), and plates were incubated at 37°C for 1 hour.
3. 7 μl of sodium acetate (0.3 M) and 40 μl of 96% ethanol was added to each sample before spinning at 3200 rpm, 20°C for 20 minutes.
4. The supernatant was then discarded, and the pellet was dried before adding 100 μl of 70% ethanol and centrifuging at 3200 rpm, 20°C for 10 minutes.
5. The supernatant was discarded and the pellet was dried and resuspended in 5 μl of TE (10:0.1).
6. Prior to loading on a 377 ABI Automated DNA sequencer, 2 μl of the fluorescent marker was added to each well and samples were denatured by placing them for 10 minutes on a block heated to 80°C.

2.2.4. Data analysis

Data from the ABI sequencer is processed by the ‘Image’ program (Production Software Group, The Sanger Centre, unpublished, based on Sulston *et al.* (1988). ‘Image’ automatically tracks lanes on a gel, distinguishes bands and normalizes bands against a marker lane, although the first two steps required checking and normally some form of manual alteration. ‘FPC’ (Soderlund *et al.*, 1997, Soderlund *et al.*, 2000) takes as input a set of clones and the bands, corresponding to restriction fragments, called by the ‘Image’ package for each clone. ‘FPC’ calculates the probability of two clones overlapping based on the similarity of their fragments, using an algorithm based on the probability of coincidence score. Contigs consisting of two or more clones can be built according to these scores, and ‘FPC’ then builds a consensus band (CB) map for each contig. The CB map is used to assign coordinates to the clones based on their position with regard to this consensus map, providing a detailed picture of how much clones overlap within the contig. Some degree of manual editing is generally required: clones can be removed and added to contigs, clone coordinates can be refined, and contigs can be merged, split or deleted.

2.3. Fluorescent *in situ* hybridisation (FISH) mapping

FISH analysis of the mouse clone, bM573K1, was performed in order to confirm that the clone could be mapped to mouse chromosome 17. This work was carried out in collaboration with the Human Cytogenetics Laboratory, ICRF, London (Denise Sheer and Jill Williamson). The protocol was adapted from Senger *et al.* (1993): it is based on labelling the probe using nick-translation.

1. 1 µg of the clone was labelled with biotin-14-dATP using the Bionick labelling system (Gibco-BRL). The reaction was incubated for 1 hour at 15°C.
2. In order to consider the efficacy of the reaction, 100ng of the nick-labelled DNA was loaded onto a 1% agarose gel, and fragments were compared against a 100 bp marker (Gibco-BRL). Fragments of 500 bp are an ideal size for FISH but a range of 200-1000 bp was considered acceptable.
3. As fragments of the required size were produced, the labelling reaction was halted by adding 5 µl of EDTA.
4. 100 ng of the labelling reaction was mixed with 2 µg human Cot-1 DNA (Gibco-BRL), 2 µl 3M sodium acetate (pH 5.6), and 50 µl 100% ethanol. This mixture was placed on dry ice for 1 hour to precipitate the probe.

5. The mixture was centrifuged at 14000 rpm, 4°C for 15 minutes and the pellet was dried in a speed vacuum, before resuspension in 11 μl of hybridisation solution.
6. The resuspended DNA was denatured at 85°C for 5 minutes. Incubation at 37°C for 30 minutes followed in order to compete out the repetitive elements.
7. The metaphase slides were denatured in 100 μl of 70% formamide, 2 x SSC (pH 7.0) at 75°C for 2.5 minutes.
8. The slides were dehydrated through exposure to 3 concentrations of ethanol: slides were shaken for 3.5 minutes with 70% ethanol, 95% ethanol and finally 100% ethanol.
9. The slides were air dried to evaporate the remaining ethanol.
10. The hybridisation mix was placed onto the denatured slides and there were covered by 20 x 20 mm cover slips. Rubber cement was used to seal the cover slips and the slides were incubated in a moist chamber at 37°C for 24 hours.
11. After hybridisation, slides were washed 3 times at 42°C in 50% formamide, 2 x SSC (pH 7.0).
12. Further washing was performed 3 times at 42°C in 2 x SSC (pH 7.0). Finally, the slide was briefly washed in SSCTM solution.
13. For detection of the biotinylated probes, the slide was incubated with a 1:500 dilution of avidin-FITC (Vector laboratories) in SSCTM solution. Incubation was at 37°C for 30 minutes.
14. The slide was counterstained with 0.06 μg/ml of DAPI (4',6'-diamidino-2-phenylindole hydrochloride) in Citifluor AF1 (an antifadent contained in a glycerol PBS which maintains fluorescence).
15. 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 (Vysis, UK).

2.4. The production of PUC and M13 shotgun libraries (Bankier *et al.*, 1987)

Using the large-scale maps produced by FPC fingerprinting methods, the minimal number of clones that would cover the region were selected for shotgun sequencing. Shotgun sequencing involves generating a random set of fragments which are then assembled so overlapping fragments of sequence provide the complete sequence across the clone. Typically, to declare a clone sequence accurate, there is a requirement for each section of the clone to be covered by 6-8 separate fragments; this redundancy allows for the resolution of sequencing errors. As part of the large-scale sequencing effort, human PACs and BACs (from the libraries RPCI1 and RPCI-11.1 constructed at the Roswell Park Cancer Institute by the group of Pieter de Jong) were subcloned at the Sanger Centre using the method described below; mouse BACs (from the Research Genetics CITB-CJ7-B library) and PACs (from library RPCI-21, also constructed at Roswell Park Cancer Institute) were supplied by Claire Amadou (Amadou *et al.*, 1999) and subcloned by me.

2.4.1. Isolation of PAC/BAC DNA using the caesium chloride procedure

1. A 500 ml sterile plugged flask containing 200 ml 2x TY and the appropriate selective agent (0.6 ml kanamycin (25 mg/ml) for PACs, 0.1 ml chloramphenicol (25 mg/ml) for BACs) was inoculated with a single colony, and incubated (with shaking at 300 rpm) at 37°C for 18-24 hours.
2. The cells and medium were then transferred to a 250 ml bottle and spun at 6000 rpm for 5 minutes (Sorvall centrifuge).
3. The supernatant was discarded, and the pellet was dried briefly by draining the bottle onto tissue, before being fully resuspended (by drawing the mixture up and down the pipette) in 50 ml GET solution.

4. 50 ml of NaOH/SDS solution was added to the bottle, which was inverted gently 3-4 times, and then left for 5 minutes at room temperature.
5. 50 ml of potassium acetate solution was added, and the bottle was inverted 10-12 times before being placed in an ice-bath for 20 minutes, and then centrifuged at 12000 rpm, 4°C for 20 minutes.
6. After centrifugation, the supernatant was filtered through a piece of sterile cheese cloth into a new 250 ml bottle.
7. 90 ml isopropanol was added to the supernatant, which was then left at room temperature for 5 minutes before spinning at 9000 rpm, 4°C for 15 minutes.
8. The supernatant was discarded and the pellet was washed with 25 ml 70% ethanol, centrifuging at 9000 rpm for 5 minutes.
9. The pellet was dried by removing all traces of the supernatant and then placing the bottle in a vacuum-drier for 2-3 hours.
10. To resuspend the pellet 2.9 ml of TE (10:1) was added and the bottle was swirled gently.
11. 3 g caesium chloride were added to a 50 ml falcon tube, into which the DNA solution was then transferred.
12. After the caesium chloride had dissolved, 290 ml ethidium bromide was added to the tube and the solution was spun at 3000 rpm, 20°C for 12 minutes before being transferred into a TL100 tube.
13. The TL100 tube was heat-sealed and centrifuged at 70000 rpm, 20°C for 16-24 hours.
14. From the TL100 tube, the lower (supercoiled) band of DNA was removed using a 1 ml syringe with a 20G needle. The amount of DNA recovered (about 200-300 μl) was placed in a 1.5 ml eppendorf tube.
15. 0.3 ml of water and 0.5 ml of isobutanol were added to the 1.5 ml eppendorf tube.

16. After mixing, this produced one immiscible (pink) layer which was discarded, leaving 0.4-0.5 ml of solution in the tube.
17. 2 volumes of ethanol (0.8-1.0 ml) was added to the tube and it was placed at 4°C for 5 minutes, before being spun at 1500 rpm for a further 5 minutes.
18. The supernatant was discarded and the pellet was resuspended in 400 µl of TE (10:0.1) by vortexing the tube, chilling at 4°C for 15 minutes, and then vortexing again.
19. 40 µl of sodium acetate/EDTA solution was added, followed by 2 volumes of ethanol (0.8-1.0 ml). Mixing followed the addition of both solutions; the tube was then placed at -20°C for at least 30 minutes.
20. The tube was spun at 1500 rpm for 5 minutes, and the supernatant was discarded.
21. 1 ml of 80% ethanol was added and the tube was centrifuged at 1500 rpm for 5 minutes. The supernatant was removed and final traces of ethanol were left to drain out of the tube for 5 minutes, before vacuum-drying for 5-10 minutes.
22. Resuspension was performed, using 20 µl of TE (10:0.1).

2.4.2. Sonication and subfragment end repair of plasmid DNA

1. In order to estimate the concentration of DNA in the BAC/PAC, a 0.5% agarose mini-gel was run on a 10 x dilution of the sample. A gel was prepared using 50 ml 1 x TBE and 0.25 g of agarose, and samples were run alongside λHindIII/pBR322 markers. Samples were visualized by soaking the gel in 500 ml of 1 x TBE containing 25 µl ethidium bromide (10 mg/ml).
2. From this gel picture, the amount of DNA required to obtain 10 µg was taken for sonication. Water was added so the total volume of water and DNA was 54 µl. 6µl of mung bean buffer was added to this and the mixture was vortexed.

3. The tube was placed in the cup horn containing ice cold water inside the sonicator (in a cold room). The tube was positioned about 1 mm away from the face of the probe.
4. 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. (Required outputs/ time vary according to the specifications of the sonicator, and the size of fragments which are desired).
5. If no movement and cavitation of the cup and tube could be observed, sonication was performed again. The mixture was briefly centrifuged at 10000 rpm.
6. 1 μl of sonicated DNA was mixed with 4 μl of loading buffer and the sample was run alongside λHindIII/pBR322 markers on a 0.8% minigel. (0.4g agarose, 50 ml 1 x TBE).
7. The DNA was checked after sonication: the ideal outcome was a smear with no sign of a band of high molecular weight DNA. Near complete sonication was also observed (a smear with a faint band of high molecular weight), and unsonicated samples showing only faint smearing with a substantial band of high molecular weight were also present.
8. Unsonicated samples were sonicated again as above. A second check gel was run to see if these samples had been fragmented. Samples showing incomplete sonication were sonicated for 5 further seconds.
9. 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.
10. The volume in the tube was made up to 200 μl with water, and 20 μl of 1 M sodium chloride, 550 μl of ice cold 100% ethanol, and 1 μl of pellet paint were added to the DNA.
11. In order to precipitate the DNA, it was left overnight (or for at least 2 hours) at -20°C and then centrifuged for 30 minutes at 4°C, 13000 rpm.

12. The supernatant was removed from the tube, leaving the DNA pellet which was washed in 1 ml 100% ethanol by centrifugation for 10 minutes at 4°C, 13000 rpm.
13. The ethanol was removed and the pellet was dried in a vacuum dryer for 10-15 minutes.

2.4.3. Selection of suitably sized DNA fragments for subcloning

1. A 0.8% TAE gel (0.4 g agarose, 50 ml 1 x TAE, 2 μl ethidium bromide) was made and was placed in a gel tank containing 500 ml of 1 x TAE and 20 μl of ethidium bromide (10 mg/ml).
2. The pellet was resuspended for loading in 6.25 μl of TE (10:0.1), 0.75 μl 10 x TAE, and 2 μl of loading dye. Care was taken to ensure all the DNA pellet was incorporated in this mixture.
3. All (9 μl) of this mix was loaded alongside λHindIII/pBR322 markers, and the gel was run at 35 mA, 50-60 V for approximately 2 hours.
4. On the long wave ultra violet transilluminator, bands corresponding to the 1.4-2 Kb (ideal) size were cut out. Additional bands of 0.6-1 Kb, 1-1.4 Kb and 2-4 Kb were also cut from the gel: these were stored in case they were needed at a later stage.
5. The pieces of gel were weighed so gel volumes could be estimated.
6. The 1.4-2 Kb gel fragment was placed in a tube and incubated at 65 °C for 5-10 minutes.
7. 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.
8. 30 μl of sodium chloride, 200 μl of buffered phenol and 196 μl (corresponding to the weight of the gel piece) of TE (10:1) buffer were added to the tube.
9. The tube was spun down for 3 minutes at 13000 rpm and the upper (aqueous) phase (about 230 μl) was extracted and added to a new tube.

10. 100 µl of TE (10:1) was added to the old mixture which was respun at 13000 rpm for 3 minutes. The upper layer (about 100 µl) was extracted and added to the 230 µl extracted earlier, whilst the organic phase was discarded.
11. 130 µl of isobutanol was added to the new tube which was spun at 13000 rpm for 1 minute. The aqueous layer was extracted and discarded.
12. 1 µl of pellet paint (Novagen) and 700 µl 100% ethanol were added to the tube which was placed at -20°C overnight (or for a minimum of 30 minutes).
13. The tube was spun at 4°C, 13000 rpm for 30 minutes, and the ethanol was decanted out of the tube.
14. The pellet was resuspended in 1 ml of ethanol and spun at 4°C, 13000 rpm for 10 minutes.
15. Ethanol was removed from the pellet, which was vacuum dried for 5-10 minutes before resuspension in 5 µl of TE (10:0.1).
16. To check for successful elution, 0.5 µl of DNA with 4.5 µl of loading dye was run out on a 0.8% TBE agarose gel with λHindIII/pBR322 markers.

2.4.4. Ligation and transformations

2.4.4.i. Ligation into pUC18 vector

1. A premix of pUC18 (SmaI/CIP, Amersham) and buffer, 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. 0.15µl of the pUC18-buffer mix was dispensed into the 0.5 ml tubes 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 water (b) 0.7 µl water (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 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 water was added to each reaction, and tubes were stored at -20°C until transformations were performed.

2.4.4.ii. Ligation into M13 vector

1. A premix consisting of 0.2 µl M13mp18 (SmaI/CIP, Amersham) per reaction and 0.2 µl buffer (supplied with the vector) was made up.
2. 0.4 µl of this mix and 1.4 µl of DNA was dispensed into each tube.
3. As with the pUC18 ligations, 3 controls were set-up: (a) 1.4 µl water (b) 1.4 µl water (c) 1.0 µl phix174/HaeIII (2.0 ng)
4. With the exception of tube (b), 0.2 µl T4 ligase was added to each tube, and tubes were shaken and centrifuged gently.
5. Tubes were transferred to a 16°C incubator overnight.
6. Tubes were heated to 65°C for 7 minutes, before being left at room temperature for 5 minutes, chilled on ice, and then centrifuged briefly.
7. 18 µl of water was added to each reaction and tubes were stored at -20°C.

2.4.4.iii. Transformations of pUC18 vectors

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 Eppendorf 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 then 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 in the Eppendorf tube.
4. The cells, glycerol and DNA were aliquoted into a cuvette placed on ice
5. The SOC solution was warmed in a water bath (20-30°C) and the solution was taken up in a Pasteur pipette.
6. The cuvette containing the DNA ands cells was placed in an electroporator, which was set to deliver a pulse in the range 3.8-5.0. (This range had been optimised by control experiments assessing the efficiency of transformation at a range of electric pulses.)
7. The cuvette was removed from the electroporator and 400 µl SOC was added to the cuvette: the mixture of SOC, cells and DNA was taken up and ejected into a test-tube.
8. Test-tubes were incubated in a shaker at 30°C for 1 hour with agitation.
9. TYE/Amp plates (90 mm) were placed at room temperature.
10. Test-tubes were removed from the shaker and 50 µl IPTG (40 mg/ml) and 50 µl Xgal (50 mg/ml) were added to each tube.
11. 125 µl of the solution was dispensed onto one TYE/Amp plate and 250 µl was dispensed onto a second plate.
12. A sterile spreader was used to make the solution cover the plate in an even manner.
13. Plates were placed in a 37°C incubator overnight.

2.4.4.iv. Transformations of M13 vectors

1. TYE/AMP plates were placed in a 37°C incubator, and 1 litre of H-Top agar was melted in a microwave.
2. 0.2 µl of ligated DNA was dispensed into a 1 ml eppendorf tube which was placed in a heated rack.
3. 3 ml of H-Top agar was added to one glass test-tube for each reaction, along with 25 µl IPTG (40 mg/ml) and 25 µl Xgal (25mg/ml).
4. Each tube of TG-1 cells was mixed with 150 µl 10% glycerol, and 40 µl of this mixture was added to the ligated DNA in the Eppendorf tube.
5. This mixture was added to a cuvette placed on ice.
6. Plates were removed from the incubator, and warmed SOC was taken up in a Pasteur pipette.
7. The cuvette was placed in the electroporator, and a pulse of 4.4-4.6 was delivered to the cells.
8. The warmed SOC (400 µl) was used to dilute the mixture in the cuvette which was then transferred to a test-tube containing the H-Top, IPTG and Xgal.
9. The contents of the test-tube was mixed by rolling the tube once between the palms.
10. The mixture was then emptied onto a TYE/Amp plate and the plate was swirled until an even coverage was obtained.
11. Once the H-Top had set, plates were placed in a 37°C incubator overnight.

2.5. Shotgun sequencing.

In the high throughput system operated at the Sanger Centre, successful ligations were stored at -20°C until clones were selected for shotgun sequencing. Upon selection, a number of plates (5 pUC, 5 M13) were produced from the ligations. Colonies from this plates were picked into 96 well plates containing the appropriate growth media, and after growth, DNA was prepared for sequencing. Prior to the preparation of this DNA, cells from each well were transferred into a 96 well plate (Corning) containing glycerol; this provided a stock of cells allowing inserts from a specific well to be regrown if necessary. These back-up stocks were useful if DNA from a specific insert was required to assembly a clone.

2.5.1. Preparation of template DNA in M13 vector (based on Mardis (1994)).

1. 5 ml of TG-1 cells was added to 500 ml of 2 xTY media containing ampicillin, and 1 ml of this solution was aliquoted into each well of a 96 well Beckman box. (Either by hand or by using an automated plate filler).
2. Colonies from TYE/Amp plates were picked into these wells. (Either by hand or using an automated picking machine.)
3. Boxes were sealed and lids were pierced for aeration, before boxes were placed in a 37°C incubator at 360 rpm for 12.5 hours.
4. 100 μl of the cells were removed from each well and added to a 96 well plate (Corning) containing 50 μl 10% glycerol. These plates were sealed and stored at -70°C.
5. Boxes were centrifuged at 4000 rpm for 2 minutes.
6. New Beckman boxes, with each well containing 145 μl of 20% PEG 8000, were set-up.
7. After centrifugation, 580 μl of the supernatant was transferred from the Beckman boxes containing cells into the Beckman boxes containing the PEG.

8. Boxes were sealed and shaken (by hand), and left for 20 minutes at room temperature.
9. Boxes were centrifuged at 4000 rpm for 20 minutes, and the supernatant was discarded with boxes being drained onto paper towels.
10. Boxes were spun upside down on towels at 300 rpm for 2 minutes to remove lingering traces of supernatant.
11. 20 µl of triton was added per well and boxes were sealed with silver foil, before being strongly vortexed, briefly spun to 1000 rpm, and vortexed and spun once more.
12. Boxes were placed in a 80°C water bath for 10 minutes.
13. Boxes were centrifuged to 1000 rpm, 40 µl water was added, and boxes were spun to 1000 rpm again.
14. 96 well microtitre plates (Serocluster), containing 170 µl EtOH/NaOAC mix per well, were set-up.
15. The contents of each well (60µl) of the Beckman box were transferred into the prepared microtitre plate, and solutions were mixed by pipetting up and down.
16. Microtitre plates were centrifuged at 4000 rpm for 60 minutes.
17. The supernatant was decanted from the plates, which were drained on towels before adding 200 µl of ice-cold 70% ethanol.
18. Plates were centrifuged at 4000 rpm for 15 minutes, the supernatant was decanted and plates were drained on towels.
19. Finally, plates were placed in a 37°C oven for 30-60 minutes in order to dry the pellets which were then resuspended in 60 µl 0.1 mM EDTA.

2.5.2. Preparation of template DNA in pUC18 vector

1. 1 ml of circlegrow containing ampicillin was aliquoted into each well of a 96 well Beckman box, and separate 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 and left to grow for 22 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 10% glycerol. These plates were sealed and stored at -70°C.
4. Boxes were spun for 5 minutes at 4000 rpm, the supernatant was discarded and boxes were placed upside down on towels for 20 minutes to dry.
5. 250 µl GET solution (Solution 1) was added to each well and cells were vortexed for 2 minutes.
6. Boxes were spun at 4000 rpm for 5 minutes to pellet cells.
7. The supernatant was discarded and boxes were left to drain, before 250 µl GET solution was added and boxes were vortexed for 2 minutes.
8. Microtitre plates (Serocluster) containing 4 µl RNase A (20 mg/ml) were set-up.
9. From each well, 60 µl of the resuspended cells were transferred to these Serocluster plates.
10. 60 µl NaOH/SDS solution was added to each well and plates were sealed with 3M plate sealers (Scotch), before solutions were mixed by inversion (10 times).
11. Plates were left at room temperature for 10 minutes.
12. 60 µl potassium acetate (3 M) was added, plates were sealed and solutions were mixed by inversion (10 times).

13. Plates were left at room temperature for 10 minutes, plate sealers were removed and plates were placed in a 90°C oven for 30 minutes.
14. The plate was placed on ice for 5 minutes.
15. Filter plates were prepared by taping a Millipore 96 well filter plate on top of a Falcon 96 well plate.
16. The contents of each well was transferred from the Serocluster plate into the filter plate, and filter plates were spun at 3300 rpm for 2 minutes.
17. The top filter plate was discarded and 110 µl isopropanol was added to the filtrate.
18. Plates were sealed and mixed by inversion (twice), before spinning at 3750 rpm for 30 minutes.
19. The supernatant was discarded and 200 µl ice-cold 70% ethanol was added to the plates which were spun for 5 minutes at 3750 rpm.
20. The supernatant was discarded and plates were allowed to drain on towels.
21. When the pellet was totally dry, it was resuspended in 35 µl of 1 M Tris-HCl, 0.1 mM EDTA.

2.5.3. The sequencing reaction.

2.5.3.i .Using dye primers:

1. 2 µl of DNA was aliquoted into 4 wells, and 8 µl of the specific dye primer ready reaction mix (ABI) was added to each well. (Each of the 4 wells should contain only one of the 4 dye primers.)
2. This mixture was spun and placed on a thermocycler with the following program:

   (i) 92°C for 15 seconds (ii) 50°C for 15 seconds (iii) 70°C for 1 minute, (iv) repeat (i) – (iii) for 20 cycles (v) 4°C until stopped.

3. The DNA from all 4 wells was combined in one well of a Serocluster plate, the plate was spun to 1000 rpm, and 10 µl 3M sodium acetate (pH 4.8) and 160 µl 96% ice-cold ethanol were added.
4. The plate was spun at 4°C, 4000 rpm for 90 minutes, and the ethanol was decanted.
5. 200 µl of 70% ice-cold ethanol was added, and the plate was spun for 15 minutes at 4°C, 4000 rpm.
6. The ethanol was removed and the pellet was dried for sequencing. Alternatively, plates were stored at -20°C until sequencing space became available.

2.5.3.ii. Using dye terminators:

1. 3 µl of DNA was added to 9 µl of a mix made up of 1 µl primer (6 pM), 4 µl dye terminator ready reaction mix (ABI) and 4 µl water.
2. The mixture was spun and placed on a thermocycler with the following program:
   (i) 96°C for 10 seconds (ii) 50°C for 5 seconds (iii) 60°C for 4 minutes, (iv) repeat (i) – (iii) for 25 cycles (v) 4°C until stopped.
3. The DNA was transferred to a Serocluster plate and 10 µl 3 M sodium acetate (pH 4.8) and 160 µl 96% ice-cold ethanol were added.
4. The plate was spun at 4°C, 4000 rpm for 60 minutes, and the ethanol was decanted.
5. Steps 5-6 as described above (for the dye primers) were performed.

2.5.4. Sequencing instrumentation.

Clones sequenced by me were loaded on either an ABIPRISM 373 sequencer or an ABIPRISM 377 sequencer (ABI, Foster City, USA).

2.5.4.i. ABI-373 set-up:

1. 3 μl of sequencer loading dye was added to each well, and samples were briefly centrifuged.
2. The gel was inserted into the machine, and after cleaning the glass plate around the laser, the machine was plate-checked: if the glass plate appeared clear then 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.
3. Samples were denatured by heating at 80°C for 10 minutes before loading.
4. The comb was removed from the gel and wells were washed out before samples (36 at most) were loaded using a Gilson pipette.
5. Data was collected over a run-time of 8 hours.

2.5.4.ii. ABI-377 set-up:

1. 2 μl of loading dye was added to each well, and samples were briefly centrifuged.
2. The gel was inserted into the machine, and after cleaning the area of the gel around the laser, the machine was plate-checked.
3. The upper buffer chamber was put in place, along with the heat plate that clipped onto the front of the gel, and the machine was pre-run for 30 minutes.
4. Buffer chambers were filled with TBE, samples were denatured as above, and wells were washed out before samples were loaded (48-60 samples) and run for 4 hours.

2.6. Data analysis of shotgun sequencing reactions and clone assembly.

After a basic level of analysis, data produced from the ABI sequencers was transferred to the UNIX system where a number of programs have been developed for the analysis of this data. The first procedure involved in analysing a 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, 'Gelminder' moves onto to call the bases. Data from each sample 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 or sequence of an unacceptably poor quality. Clipped good quality sequences are then passed into the 'Phrap2Gap' program (Mott and Dear, unpublished). This program is a modified version of 'Phred' and 'Phrap' (Gordon *et al.*, 1998), which are base calling programs and sequence assembly programs respectively. '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 *et al.*, 2000, Staden *et al.*, 2001); 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' and 'GAP4.new' packages (The human clone AL031983 and the mouse clone AL078630 were both assembled by me using this software).

Generally, upon transfer of clone DNA into a 'GAP' package, the clone was not a contiguous piece of sequence 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 all the 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.

2.6.1. PCR reaction used in clone assembly

1. 1 µl (40 nM) of forward primer and 1 µl (40 nM) reverse primer were dispensed into a 96 well plate (Costar), spun briefly and dried down in a 90°C oven.
2. A master mix was made: per sample, 5 µl PCR buffer (AmpliTaq), 2 µl 4 x dNTPs, 2 µl AmpliTaq, 33 µl water.
3. 2 µl of the appropriate DNA (taken from DNA stock plates) was added to the well, along with 47 µl of the master mix.
4. Samples were placed on the thermocycler with the following program: (i) 94°C for 1 minute, (ii) 55°C for 1 minute, (iii) 72°C for 3 minutes, (iv) steps (i)-(iii) repeated 25 times, (v) 4°C until program stopped.
5. 50 µl MgCl PEG was added to each sample, and samples were well mixed.
6. Plates were sealed and left at -20°C for 1 hour.
7. Samples were spun for 1 hour at 4°C, 4000 rpm, and the supernatant was discarded.
8. Plates were spun upside on tissue for 2 minutes at 250 rpm and 50 µl of ice-cold 96% ethanol was added to the samples.
9. Ethanol was discarded and plates were spun upside down on tissue for 2 minutes at 250 rpm.
10. After drying, 50 µl water was added to resuspend the DNA pellet.
11. Sequencing protocols were performed as above, using the appropriate primer(s).

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 below) and generating the virtual digests. Virtual digests were generated by the program ‘Confirm’ (Production Software Group, The Sanger Centre, unpublished) which also has a graphical display showing the real and virtual digests alongside each other.

2.6.2. Restriction digests used to check veracity of clone assembly

1. 30-40 ng DNA, estimated from the gel run prior to subcloning, was diluted to make a total volume of 2.5 µl.
2. 2.5 µl DNA + water, 2.5 µl of the restriction enzyme buffer (supplied with the enzyme) and 0.3 µl of the restriction enzyme (BAMHI, EcoRI, or HindIII) were placed in a well in a Costar 96 well plate.
3. The plate was spun to 1000 rpm and placed on a thermocycler at 37°C for 2 hours.
4. Reactions were placed briefly at -20°C, and then in the oven set at 60°C for 10 minutes.
5. Samples were mixed with 1 µl Ficoll dye and 7 µl of the mixture was run on a 0.7% agarose gel (1 x TAE, 200 ml gel) with 6 µl of 1 Kb ladder marker and 6 µl of λBst11071 ladder marker.
6. This gel was run overnight before staining with Vistra Green (Vistra systems): 5 ml 1 M Tris-HCl, 500 µl 0.1 M EDTA and 50 µl Vistra Green were mixed before being added to 500 ml 1 x TAE in a gel-staining tank.
7. The gel was placed in the tank containing the stain which was sealed and gently shaken for 1 hour.

2.7. Construction of mouse filters for hybridisation

In an attempt to extend the mouse contig, a number of membrane filters spotted with the clones screened during fluorescent fingerprinting were produced.

1. Fresh LB plates (8 x 12 cm, rectangular) containing the appropriate anitibiotic (kanamycin for PACs and chloramphenicol for BACs) were set-up.
2. A 8 x 12 cm piece of nylon membrane filter (Hybond N+, Amersham) was labelled with a permanent pen, and was gently placed on the surface of the agar avoiding trapping any air bubbles between the filter and the agar surface.
3. Glycerol stocks of the clones to be plated were allowed to thaw, whilst the robotic gridding system was set-up with stations containing 95% ethanol bath for sterilization, a sonication bath containing water and 0.1% Decon disinfectant, and a bath of indelible ink (autoclaved 1% Higgins black ink).
4. The thawed glycerol plate and the agar plate with the filter were placed at the appropriate positions, and the sterile ink was spotted in the pattern programmed into the robotic system.
5. Pins were cleaned in the sonication bath for 1 minute.
6. A gridding cycle, consisting of immersing the pins in ethanol for 10 seconds, followed by air drying for 10 seconds and then inoculating the culture onto the filter, was performed.
7. This step was repeated until all the clones were plated out on top of the ink spots generated in step 4; the pins were then sonicated for 1 minute.
8. Plates were incubated upside down at 37°C for 12-16 hours.
9. After 12-16 hours growth should be circular and generally uniform in size across the array: if this was achieved the next step was the lysis of the bacterial colonies.

10. Using forceps, membrane filters were removed from the agar surface, and these were placed colony-side up onto Whatman 3MM paper saturated with 10% SDS for 4 minutes.
11. Membrane filters were then placed colony-side up onto Whatman 3MM paper saturated with denaturing solution for 10 minutes.
12. Membrane filters were placed colony-side up onto Whatman 3MM paper, and allowed to air-dry for 10-20 minutes.
13. Filters were submerged in an excess of 10 x neutralizing solution for 5 minutes, with intermittent agitation. (Repeated separately for each filter).
14. Filters were submerged in an excess of 1 x neutralizing solution for 5 minutes, with intermittent agitation.
15. Filters were submerged in 2 x SSC/0.1% SDS wash solution for 5 minutes, agitating intermittently.
16. Filters were submerged in an excess of 2 x SSC for 5 minutes, again with some agitation.
17. Membrane filters were placed in an excess of 50 mM Tris-Cl and agitated intermittently (Repeated separately for each filter).
18. Membrane filters were placed colony-side up onto Whatman 3MM paper, and allowed to air-dry. These filters can be stored at room temperature for several years.
19. Prior to hybridization the DNA was cross-linked to the filters. (Amount of time for cross-linking calibrated by using a control repeated hybridisation with filters stripped for different lengths of time.)

2.8. Construction of mouse olfactory receptor gene vectors for *in situ* hybridisations

11 mouse OR genes were amplified by PCR and cloned into the pGEM T-Easy vector (Promega) system so these vectors could be used in *in situ* hybridisations of rat and mouse tissue. The pGEM

T-Easy vector system is advantageous for the cloning of PCR products, since vectors are prepared by cutting with EcoRV and adding a 3’ terminal thymidine to both ends.

The amount of PCR product to be used was calculated as follows:

$$\frac{\text{ng of vector*kb size of insert}}{\text{kb size of vector}} * \frac{\text{insert}}{\text{vector}} \text{ (ratio)}$$

2.8.1. pGEM T-Easy ligations and transformations

1. The pGEM-T Easy vector, control insert DNA and PCR products were centrifuged, and the rapid ligation buffer was vortexed.
2. The reactions were set up in 0.5 ml tubes, as follows: 5 µl 2x rapid ligation buffer, T4 ligase; 1 µl (50 ng) pGEM-T Easy vector; 1 µl (3 U) T4 DNA ligase; 36 ng of PCR product in 3 µl of water / 2 µl control insert and 1 µl water (positive control) / 3 µl water (negative control).
3. The reactions were mixed by pipetting, and were incubated overnight at 4°C.
4. Tubes containing the ligation reactions were centrifuged, and 2 µl of each reaction was transferred to a sterile 1.5 ml tube on ice. Another sterile tube, containing 0.1 ng of an uncut plasmid (pGEM) was also set-up.
5. Tubes of JM109 High Efficiency Competent cells (Promega) were thawed in an ice bath (for about 5 minutes), and mixed by gently flicking the tube.
6. 50 µl of cells were transferred into each 1.5 ml tube, the contents were mixed by flicking the tubes and tubes were then placed on ice for 20 minutes.
7. Cells were heat-shocked for 45-50 seconds in a water bath at exactly 42°C, before tubes were returned to ice for 2 minutes.

8. 950 µl of SOC medium was added to tubes containing cells transformed with ligation reactions, and 900 µl was added to the tube containing the uncut plasmid.
9. Tubes were then shaken and incubated for 1.5 hours at 37°C (150 rpm).
10. 100 µl of each culture was plated out onto 2 LB/ampicillin/IPTG/X-Gal plates.
11. Plates were incubated for 16-24 hours at 37°C.

2.8.2. pGEM T-Easy Preparation of DNA

1. Colonies were picked and used to inoculate 3 ml of LB containing ampicillin. The broth was then incubated overnight at 37°C.
2. Tubes were centrifuged for 5 minutes at 10000 rpm to pellet bacteria and the medium was discarded.
3. 250 µl of cell resuspension solution was added and the pellet was resuspended by vortexing or pipetting, before resuspended cells were transferred to a sterile 1.5 ml tube.
4. 250 µl cell lysis solution was added and the tube was inverted 4 times. The mixture was incubated at room temperature until the cell solution cleared (1-5 minutes).
5. 10 µl of alkaline protease solution was added and the tube was inverted 4 times, before incubation for 5 minutes.
6. 350 µl of neutralization solution was added and the tube was again inverted 4 times.
7. The lysate was centrifuged at 14000 rpm for 10 minutes.
8. The cleared lysate (~850µl) was transferred to the spin column (avoiding the white precipitate).
9. The supernatant was centrifuged at 14000 rpm, left for 1 minute at room temperature, and the flowthrough was discarded.
10. 750 µl column wash solution (diluted with 95% ethanol) was added.

11. The tube was centrifuged at 14000 rpm for 1 minute and the flowthrough was discarded.
12. 250 µl column wash solution was added and the tube was centrifuged at 14000 rpm for 2 minutes.
13. The spin column was transferred to a new tube, and 100 µl of nuclease free water was added before centrifugation at 14000 rpm for 1 minute.
14. The DNA was precipitated by adding 50 µl of 7.5 M ammonium acetate, 375 µl 95% ethanol, and centrifuging the sample at 14000 rpm for 15 minutes.
15. The pellet was washed in 250 µl 70% ethanol and centrifuged at 14000 rpm for 5 minutes.
16. The pellet was dried and resuspended in 10-25µl nuclease free water, before storage at -20°C.

2.9. Construction of ‘olfactory promoter region’ pGL3 luciferase reporter vectors

Having found a putative promoter region for the MHC-linked olfactory receptor gene cluster, this region was cloned into the pGL3 luciferase reporter vector (Promega). This vector allows inserts to be analysed for their ability to regulate mammalian gene expression. The pGL3 vector contains a modified coding region for firefly (*Photinus pyralis*) luciferase that has been optimised for monitoring transcriptional activity in transfected eukaryotic cells. Together with the pGL3 basic vector, the pGL3 control vector (containing promoter and enhancer), and the pGL3 promoter vector (containing the promoter but no enhancer), constructed reporter vectors were transfected into Odora and HEK293 cell-lines using the the SuperFect transfection reagent (Qiagen).

2.9.1. pGL3 reporter vector restriction digests, ligations and transformations

BglII digestion: 20 µl DNA (50 ng/µl), 20 µl 'red' buffer (ABgene), 1 µl BglII (ABgene), 9 µl water. Digested at 37°C for 2 hours.

BglII/XhoI digestion: 10 µl DNA(50 ng/µl), 20 µl 'red' buffer (ABgene), 1 µl BglII (ABgene), 1 µl XhoI (ABgene), 18 µl water. Digested at 37°C for 2 hours.

1. Tubes containing 2 µl pGL3 reporter vector DNA (50 ng/µl), 1 µl ligase buffer, 0.5 µl T4 DNA ligase, and 6.5 µl of water containing approximately 40 ng of purified PCR product were set-up.
2. The reaction was incubated for 3 hours at room temperature.
3. 5 µl of each reaction was transferred to a sterile 1.5 ml tube on ice, and tubes of JM109 High Efficiency Competent cells (Promega) were thawed in an ice bath (for about 5 minutes), and mixed by gently flicking the tube.
4. 50 µl of cells were transferred into each 1.5 ml tube, the contents were mixed by flicking the tubes and tubes were then placed on ice for 20 minutes.
5. Cells were heat-shocked for 45-50 seconds in a water bath at 42°C, and tubes were returned to ice for 2 minutes.
6. 950 µl of SOC medium was added to tubes containing cells transformed with ligation reactions, and tubes were then shaken and incubated for 1.5 hours at 37°C (150 rpm).
7. 100 µl of each culture was plated out onto 2 LB/ampicillin/IPTG/X-Gal plates which were incubated for 16-24 hours at 37°C.
8. Colonies were picked into 1 ml of LB broth containing ampicillin and grown overnight at 37°C.

2.9.2. Preparation and sequencing of pGL3 reporter vector DNA

1. Cultures were prepped using the plasmid miniprep purification kit (Qiagen), and the DNA pellet was eluted in 30 µl TE buffer.
2. A 0.8% check gel was loaded with 4 µl of the sample and 4 µl loading dye and run against a 4 µl sample of the unligated vector.
3. A number of samples from each type of PCR product (forward / reverse) were sequenced (see earlier section) in order to check the ligation had been successful.

2.9.3. Transfections of pGL3 reporter vector DNA

HEK293 (Graham *et al.*, 1977) and Odora cell (Murrell and Hunter, 1999) growth medium (with serum, protein and antibiotics): 500 ml Dulbecco's Modified Eagle Medium (DMEM high glucose without L-glutamine and sodium pyruvate, GIBCO), 10% fetal calf serum (FCS, GIBCO), 5 ml 100x L-glutamine, 1 ml 100x penicillin/streptomycin. Stored at 4°C.

1. A cell culture with 70-80% confluence was taken, the growth medium was aspirated using a Pasteur pipette, and the cells were washed twice with 10 ml PBS.
2. 4 ml trypsin-EDTA was added to the cell plate and the plate was swirled gently until cells were detached (for approximately 4 minutes).
3. 6 ml growth medium was placed in a 50 ml tube (Falcon) and the cell suspension was transferred to the tube and mixed with the medium.
4. Cells were centrifuged at 1000 rpm for 5 minutes.
5. The supernatant was removed with a Pasteur pipette and cells were washed with PBS before being resuspended in 10 ml growth medium.

6. 250 µl , 500 µl and 1 ml cell suspension were added to 3 100mm cell culture plates filled with 10 ml growth medium.
7. Plates were swirled gently and put in a 37°C incubator, 5% carbon dioxide for 2 days.
8. Cells were counted using a haemocytometer.
9. 5 x 10^5 cells were seeded in 5 ml growth medium (DMEM + serum, protein and antibiotics) in 60 mm dishes.
10. Cells were incubated at 37°C and 5% carbon dioxide for 2 days (until they had reached 40-80% confluence).
11. 5 µg vector DNA (contained in no more than 50 µl of TE buffer, ie. transfection required a concentration of greater than 0.1 µg/µl) was placed in a 5 ml tube (Eppendorf) with cell growth medium (DMEM containing no serum, proteins or antibiotics) making the total volume up to 150 µl.
12. 20 µl SuperFect transfection reagent was added to the tube, and solutions were mixed by pipetting up and down 5 times.
13. Tubes were incubated at 5-10 minutes room temperature to allow transfection complex formation.
14. The growth medium from cells in the 60 mm dishes was aspirated and cells were washed once with 4 ml PBS.
15. 1 ml cell growth medium (DMEM containing serum and antibiotics) was added to the 5 ml tube containing the vector DNA.
16. Solutions were mixed by pipetting up and down twice, before the total volume was transferred to the cells in the 60 mm dishes.
17. The cells were incubated for 2-3 hours at 37°C, 5% carbon dioxide.
18. The medium was removed by aspiration and the cells were washed 4 times in 4 x PBS.

19. Fresh cell growth medium (DMEM with serum and antibiotics) was added and cells were incubated for 48 hours.

2.10. The pGL3 reporter vector assay

Luciferase activity was determined on the cells in medium with Bright-Glo reagent (Promega). This reagent contains beetle luciferin that is coverted into oxyluciferin by firefly luciferase, releasing a large amount of light energy that can be measured using a luminometer. The Bright-Glo Luciferase Assay system has been designed for use with cells in their growth medium.

1. Cells were counted using a haemocytometer, and the amount of medium containing 2 x 10^4 cells was transferred to a 1.5 ml Eppendorf tube.
2. The volume in the tube was made up to 100 µl using growth medium, and cells were left to equilibrate to room temperature.
3. The Bright-Glo reagent was prepared by transferring the contents of one bottle of Bright-Glo Buffer into one bottle of Bright-Glo substrate, and mixing the solutions by inversion until the substrate was fully dissolved.
4. For each sample, a measurement of luminometer (Sirius) activity was taken prior to the addition of the reagent to control for background luminescence and differences in transparency of tubes and media in each sample.
5. For each sample, the 100 µl of cells was combined with 100 µl of Bright-Glo reagent (Promega), and after 2 minutes to allow cell lysis, a second luminometer measurement was taken.

6. Measurements of background luminescence were subtracted from the second measure of luminescence, and activity values were normalized to the average activity of the pGL3 control vector (which contains both Promoter and Enhancer sequences).
7. For each reaction, 2 separate experiments were performed (new transfections and new luminometer readings.)

2.11. Polymorphism analysis

Cell lines were derived from different donors, representing different HLA haplotypes and different ethnic origins. Eight of the 10 cell lines were HLA homozygous, whereas two (BM19.7, BM28.7) were HLA hemizygous (Ziegler *et al.*, 1985, Volz *et al.*, 1992). All cell lines were grown in RPMI 1640 medium containing antibiotics and 10% fetal calf serum.

1. For each gene, 6 primers were designed (Appendix1) and the appropriate PCRs were set-up (C→D, C→F, C→H, E→D, E→H, G→D) as follows: 2 µl genomic DNA (200 ng), 2 µl primer 1 (50 pmol), 2 µl primer 2 (50 pmol), 4 µl dNTPs (2.5 mMol/1000 ml), 5 µl 10 x PCR buffer (AmpliTaq), 0.5 µl AmpliTaq, 34.5 µl water.
2. Reactions were placed on a thermocycler with the following program: (i) 95°C for 1 minute, (ii) 94°C for 30 seconds, (iii) annealing temperature for 45 seconds, (iv) 72°C for 3 minutes, (v) repeat (ii) – (iv) 45 times, (vi) 72°C for 5 minutes, (vii) 4°C until stopped.
3. 4 µl of the reaction mixture was run out on a 0.8% agarose gel with a 100 bp ladder marker: samples containing DNA products of the correct size were then purified using the PCR purification kit (Qiagen).
4. Reactions were sequenced with the appropriate PCR primers using the protocol described earlier.

5. Sequence from these reactions was assembled in a 'GAP4' database with several reads covering each region of the gene.

2.12. Mouse RT-PCR.

Expression of OR genes in various mouse tissues and organs was investigated using RT-PCR. This was performed using unique primers from the 3' end of the gene.

1. The master-mix was made up as follows: 7.2 µl 34.6% sucrose, 0.187 µl 1 in 10 fresh β-mercaptoethanol (in TE (10:0.1)), 1 µl 10 mM dNTP, 2 µl 10 x Rxn buffer, 3.49 µl dilution buffer, 0.125 µl AmpliTaq polymerase (Perkin Elmer).
2. 2 µl of a mix of both primers (at 50 ng/µl of each primer) and 5 µl of cDNA were aliquoted into the well of a Costar 96 well plate.
3. 14 µl of the master mix was dispensed into each well, and the plate was placed on a thermocycler with the following program: (i) 92°C for 2 minutes, (ii) 92°C for 30 seconds, (iii) 55 / 57.5 / 60 °C for 90 seconds, (iv) 72°C for 1 minute, (v) Repeat (ii) – (iv) 35 or 45 times, (vi) 72°C for 10 minutes, (vii) 4°C until stopped.
4. Prescreening was done at a range of temperatures with either 35 or 45 cycles.
5. Samples were run on a 2.5% agarose gel in 1 x TBE at 200 mA for 40 minutes.

2.13. Human RNA dot-blot hybridisations.

In order to consider whether expression of olfactory receptors was found outside the olfactory epithelium, a human RNA dot-blot (Clontech) was hybridised with probes generated from the 3' end of the olfactory receptor.

1.15 ml of ExpressHyb (Clontech) was prewarmed in a water bath at 50°C and 150 μl sheared salmon testes DNA (1.5 mg) was heated at 95-100°C for 5 minutes before being chilled quickly on ice.

1. The heat-denatured salmon testes DNA was mixed with the prewarmed ExpressHyb.
2. The master blot was placed in the hybridisation container and 10 ml of the ExpressHyb and salmon testes DNA was added to the container.
3. The master blot was prehybridized for (at least) 30 minutes with continuous agitation at 65°C.
4. The probe was labelled using the High Prime Kit (Boehringer-Mannheim):
   i. 25 ng DNA in 11 μl water was combined on ice with 4 μl High Prime, and 5 μl (50μCi) [α^{32}P] dCTP, and the mixture was incubated at 37°C for 10 minutes.
   ii. Free nucleotides were removed using the QiaQuick Nucletide Removal Kit (Qiagen) as follows:
   iii. 200 μl of buffer PN was added to the reaction which was transferred to a spin column and DNA was bound by centrifuging at 6000 rpm for 1 minute.
   iv. The flowthrough was discarded and 500 μl buffer PE was added, and the tube was centrifuged at 6000 rpm for 1 minute.
   v. Step (iv) was repeated again, and the flowthrough was discarded, before centrifugation at 13000 rpm for 1 minute.
   vi. DNA was eluted with 100 μl buffer EB and by centrifugation at 13000 rpm.
5. 30 μl (30 μg) human CotI DNA, 15 μl salmon sperm DNA, 50 μl 20 x SSC and 5 μl of water was added to the DNA probe.
6. The probe mixture was added to the remaining 5 ml of ExpressHyb and salmon testes DNA solution.

7. The pre-hybridization solution was discarded from the container with the master blot and the solution containing the probe was added to this container.
8. The blot was hybridised with constant agitation at 65°C overnight.
9. Six washes were performed: the first 4 at 65°C using 2 x SSC, 1 % SDS wash solution with the last two at 55°C, using 0.1 x SSC, 0.5% SDS wash solution.
10. The blot was wrapped in a heat sealed plastic bag and exposed to an X-ray film (Fuji) for 2 days at room temperature.

2.14. Titering the olfactory epithelium phage library

1. The bacterial glycerol stock (XL1-Blue MRF' strain) were streaked onto LB-tetracycline agar plates and the plates were incubated overnight at 37°C.
2. Tubes containing 10 ml LB-tetracycline medium supplemented with 10 mM $MgSO_4$ and 0.2% (w/v) maltose were inoculated with a single colony, and grown at 37°C for 4-6 hours (with agitation), or overnight at 30°C (with agitation).
3. Cells were spun at 500 x g for 10 minutes and the supernatant was discarded.
4. Cells were gently resuspended in 5 ml sterile 10 mM $MgSO_4$.
5. Cells were diluted to an OD_{600} of 0.5 with sterile 10 mM $MgSO_4$ and the bacteria were used immediately.
6. 500 µl of the phage library was diluted in 500 µl of SM buffer, and a series of dilutions were set-up and added to the host bacteria in 15 ml tubes (Falcon): (i) 1 µl phage/SM + 200 µl host cells, (ii) 1 µl of 1:10 phage/SM:SM dilution + 200 µl host cells (iii) 1 µl of 1:100 phage/SM:SM dilution + 200 µl host cells, (iv) 1 µl of 1:1000 phage/SM:SM dilution + 200 µl host cells.

7. The phage and bacteria were incubated for 15 minutes at 37°C to allow the phage to attach to cells.
8. The following was added to each tube (15 ml Falcon tube): 2-3 ml of NZY top agar (melted and cooled to ~48°C), 15µl of 0.5 M IPTG, 50µl of X-gal [250 mg/ml (in DMF)]
9. After adding the cooled NZY top agar, the mixture was plated immediately onto NZY agar plates and the plates were allowed to set for 10 minutes and plates were inverted and then incubated overnight at 37°C.
10. Around 1 x 10^5 pfu/µg of background plaques (blue) should be produced by this procedure, with the number of recombinant plaques (white) 10-100 fold higher than the background.

2.15. PCR amplification of the olfactory epithelium phage library

1. The PCR reaction was set-up as follows:0.5 µl of the phage, 2 µl (2.5µM) primer 1, 2 µl (2.5µM) primer 2, 4 µl dNTPS (2.5 mMol/1000 ml), 5 µl 10x PCR buffer, 0.5 µl AmpliTaq, 36 µl water.
2. The thermocycler program was as follows (i) 95°C for 1 minute, (ii) 94°C for 30 seconds, (iii) annealing temperature for 45 seconds, (iv) 72°C for 3 minutes, (v) repeat (ii) – (iv) 45 times, (vi) 72°C for 5 minutes (vii) 4°C until stopped.
3. PCR was set-up using pooled phage suspensions set up in the following way:
    a. The host bacteria were grown as described above.
    b. 1 ml aliquots of bacteria and medium were dispensed into 50 ml tubes, and 1 x 10^5 pfu of phage were diluted in 1 ml of SM buffer.
    c. After 15 minutes in a 37°C waterbath, 18 ml of LB broth containing 10 mM Mg SO_4 was added to the tube.

d. 1 ml of this mixture was aliquoted into a 96 well Beckman box, the box was sealed and incubated at 37°C for 5-6 hours.

e. PCR was performed on 0.5 µl of each of the phage suspensions diluted in 0.5 µl water as follows:

   2µl (2.5µM) primer 1, 2µl (2.5µM) primer 2, 4 µl dNTPS (2.5 mMol/1000 ml), 5 µl 10x PCR buffer, 0.5 µl AmpliTaq, 35.5 µl water.

   Thermocycler program: (i) 95°C for 1 minute, (ii) 94°C for 30 seconds, (iii) annealing temperature for 45 seconds, (iv) 72°C for 3 minutes, (v) repeat (ii) – (iv) 45 times, (vi) 72°C for 5 minutes

2.16. *In silico* analysis: gene finding programs

After assembly, the clone DNA was analysed using a variety of programs. Two major methods for doing this were the use of 'NIX' at the *HGMP* and the use of 'AceDB' databases at the Sanger Centre. 'NIX' runs the sequence through a suite of programs. The initial step involves disregarding the repeats in the sequence through running the sequence through '*RepeatMasker'* (Smit and Green). This program screens DNA sequences against a library of interspersed repeats and low complexity DNA sequence (*Repbase*, (Jurka, 2000)). A file produced by 'RepeatMasker' is then run through a number of gene finding programs, including '*Grail',' Genefinder', 'Genemark', 'Fex', 'Hexon',* and *'Fgene'*. 'NIX' also performs 'BLAST' searches against a number of databases: TrEMBL, Swissprot, Unigene, mRNA, EST, EMBL, HTG, GSS, STS, Ecoli, and Vector. 'BLAST' is the acronym for the basic local alignment search tool: this 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 (Altschul *et al.*, 1990). Databases searched by 'NIX' are generally maintained by *EMBL-EBI* or *NCBI*. 'NIX' (Figure 2.1) produces a graphical output of the information produced by these programs which can be used to provide a guide to features contained within the DNA clone. Similar types of programs are used by the 'AceDB' analysis package at the Sanger Centre (Figure 2.2). One additional program used for gene prediction is '*GENSCAN*', another prediction program which predicts gene structures based on profiles of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions (Burge and Karlin, 1997, Burge and Karlin, 1998).

Using a combination of gene finding programs, together with results from 'BLAST' searches is the best way of searching for genes, since predictions from any gene finding program taken in isolation are likely to be lacking in both sensitivity and specificity (Guigo *et al.*, 2000). Gene finding programs also tend to struggle with predicting smaller exons, which was sometimes a problem with predicting olfactory receptor genes, and another limitation of these programs is that they often predict incorrect joins between predicted exons (as can be seen with the OR genes in Figure 2.1 and Figure 2.2).

2.17. Identification of genes

Genes were identified in the majority of cases by homology to proteins already present in the public databases. Identification of the start and stop positions was performed by considering the open reading frame, where this existed. In the case of pseudogenes, the start and end of the gene tended to be defined by where the similarity with other proteins started and finished rather than any open reading frame. A number of novel genes were also defined: these showed matches to unidentified proteins, cDNAs or ESTs.

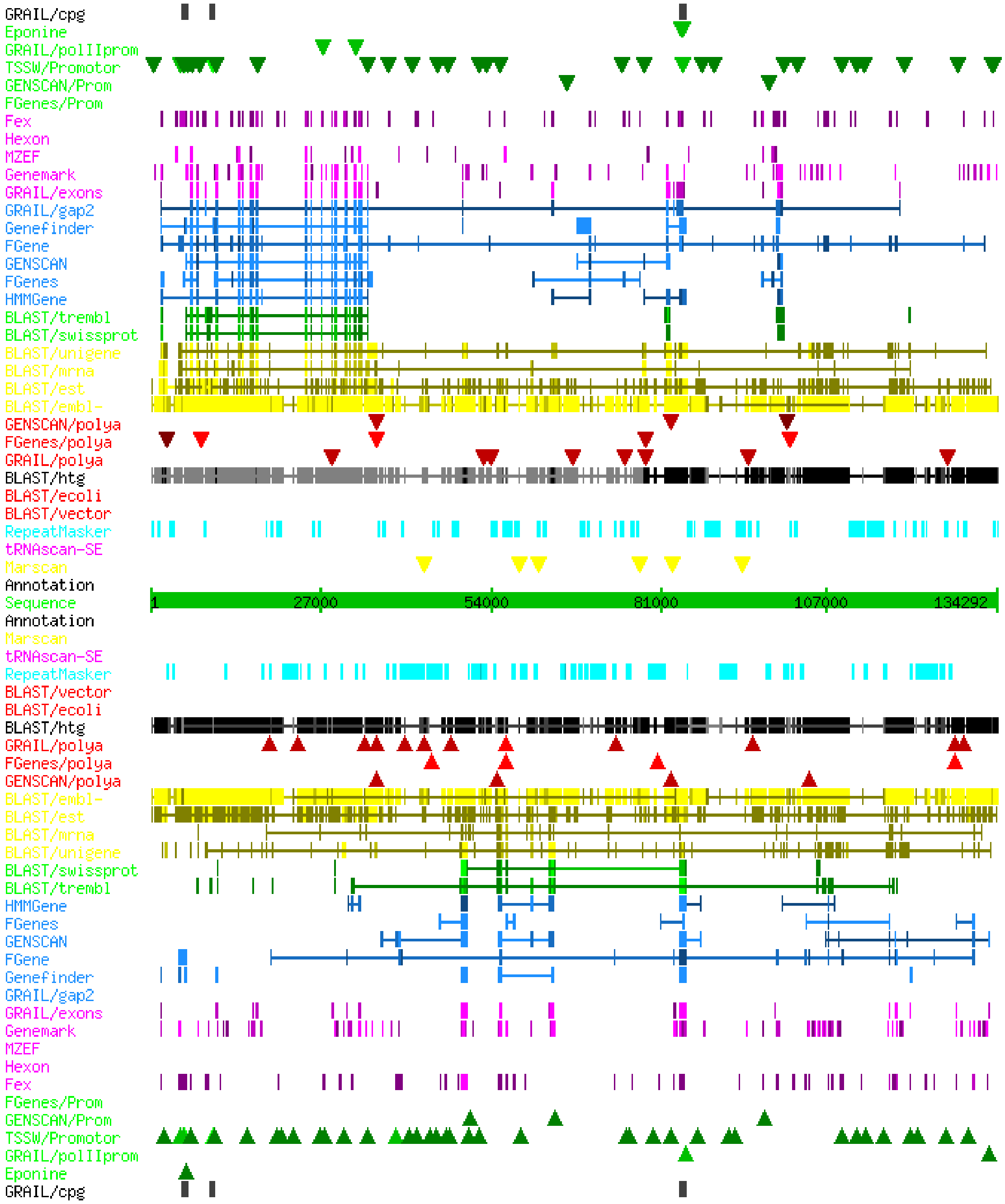
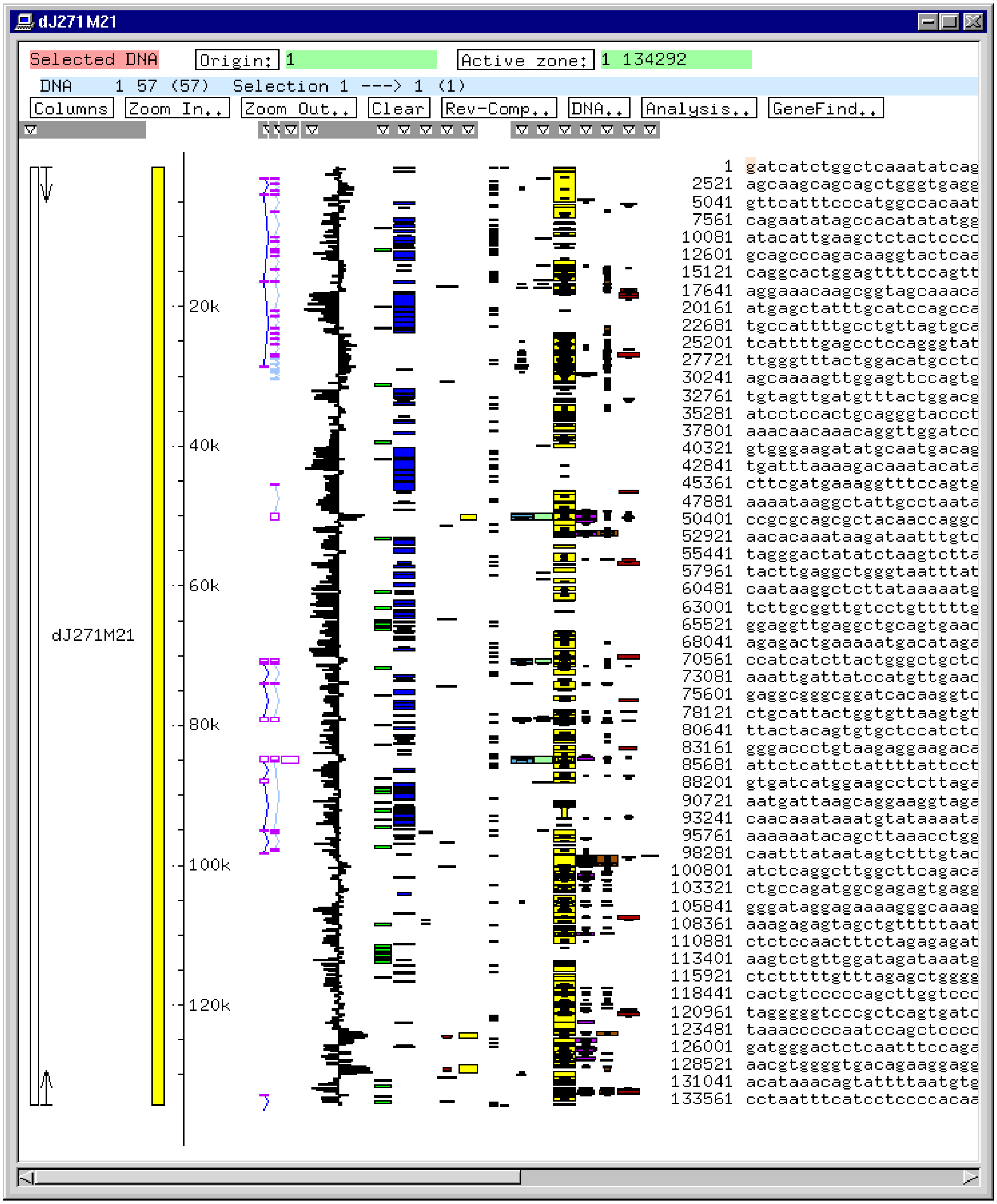


Figure 2.1: 'NIX' display of AL031983. Analyses are run on both the forward (above the green sequence line) and reverse strands (below the green line) of the clone. The column to the left shows the name of all the programs run on this piece on sequence: results appear as boxes or triangles on the line corresponding to a particular program.



```
Selected DNA   Origin: 1   Active zone: 1 134292
 DNA     1 57 (57)   Selection 1 ---> 1 (1)

     1 gatcatctggctcaaatatcag
  2521 agcaagcagcagctgggtgagg
  5041 gttcatttcccatggccacaat
  7561 cagaatatagccacatatatgg
 10081 atacattgaagctctactcccc
 12601 gcagcccagacaaggtactcaa
 15121 caggcactggagttttccagtt
 17641 aggaaacaagcggtagcaaaca
 20161 atgagctatttgcatccagcca
 22681 tgccattttgcctgttagtgca
 25201 tcattttgagcctccagggtat
 27721 ttgggtttactggacatgcctc
 30241 agcaaaagttggagttccagtg
 32761 tgtagttgatgtttactggacg
 35281 atcctccactgcagggtaccct
 37801 aaacaacaaacaggttggatcc
 40321 gtgggaagatatgcaatgacag
 42841 tgatttaaaagacaaatacata
 45361 cttcgatgaaaggtttccagtg
 47881 aaaataaggctattgcctaata
 50401 ccgcgcagcgctacaaccaggc
 52921 aacacaaataagataatttgtc
 55441 tagggactatatctaagtctta
 57961 tacttgaggctgggtaatttat
 60481 caataaggctcttataaaaatg
 63001 tcttgcggttgtcctgttttttg
 65521 ggaggttgaggctgcagtgaac
 68041 agagactgaaaaatgacataga
 70561 ccatcatcttactgggctgctc
 73081 aaattgattatccatgttgaac
 75601 gaggcgggcggatcacaaggtc
 78121 ctgcattactggtgttaagtgt
 80641 ttactacagtgtgctccatctc
 83161 gggaccctgtaagaggaagaca
 85681 attctcattctattttattcct
 88201 gtgatcatggaagcctcttaga
 90721 aatgattaagcaggaaggtaga
 93241 caacaaataaatgtataaaata
 95761 aaaaaatacagcttaaacctgg
 98281 caatttataatagtctttgtac
100801 atctcaggcttggcttcagaca
103321 ctgccagatggcgagagtgagg
105841 gggataggagaaaagggcaaag
108361 aaagagagtagctgtttttaat
110881 ctctccaactttctagagagat
113401 aagtctgttggatagataaatg
115921 ctctttttgtttagagctgggg
118441 cactgtcccccagcttggtccc
120961 taggggggtcccgctcagtgatc
123481 taaaccccaatccagctcccc
126001 gatgggactctcaatttccaga
128521 aacgtggggtgacagaaggagg
131041 acataaacagtattttaatgtg
133561 cctaatttcatcctccccacaa
```

Figure 2.2: ‘AceDB’ display of AL031983. Columns to the left of the scale show the following (left to right). Fgenes predictions, GENSCAN predictions, GC content, RepeatMasker SINEs, RepeatMasker LINEs, RepeatMasker inverted repeats, RepeatMasker tandem repeats, CpG islands, BLASTX (SwissProt) matches, GRAIL1.3 predictions, BLASTN matches (EMBL), EST matches, GSS/STS matches and polyA signals.

2.18. Identification of olfactory receptor genes

Olfactory receptor genes were defined as genes with a coding region of approximately 1 Kb. Early work produced the idea that five key protein motifs could be identified as something shared by all olfactory receptor genes: these were (i) FILLG (ii) LHTPMYFFLSHLS (iii) MAYDRYVAIC (iv) KAFSTCGSHLSVV and (v) MLNPFIYSLRN. These motifs were refined further as work progressed. Olfactory receptor pseudogenes (P), meanwhile, were defined as loci with a large number of these motifs still intact in the correct order within an approximate 1 Kb region. In contrast to functional genes, however, pseudogenes were distinguished by stops and/or frame shifts that disrupted the coding region of the gene. Genes lacking a methionine upstream of the FILLG motif (or variation) were also classified as pseudogenes, as were genes lacking any of the well-conserved motifs. Another class of olfactory receptor defined was ‘pseudogene fragments’ (PF). These loci were distinguished as having some of the conserved protein motifs, but they were generally less than 700 bp in length, and did not have 2-3 of the core motifs. Finally, a fourth class of olfactory receptor genes was the ‘fragment’ class (F): these appeared to be functional but disappeared into gaps that were present in the unfinished sequence.

2.19. The assembly of DNA clone sequences

Overlaps between clones were found using two programs, ‘cross_match’ (Green, unpublished) and ‘Dotter’ (Sonnhammer and Durbin, 1995). ‘Cross_match’ utilizes the Smith-Waterman sequence alignment algorithm (Smith and Waterman, 1981) to look for regions of similar sequence. ‘Dotter’, meanwhile, is a graphical dotplot program for detailed comparison of two sequences that compares every residue in one sequence against every residue in the other sequence. The two sequences are plotted on the X and Y axis and high scores are indicated by dots at the appropriate position.

2.20. Storage of information in sequence feature files

Information about the gene content, repeat content, and miscellaneous features (such as CpG islands, promoter predictions) was stored in files formatted so they could be used to generate postscript plots of the area. The format of these files is shown in Appendix 2. Files were either generated manually (".genes" file), or generated through a number of scripts that were developed to produce these files, for example, 'seeclone' which generated these type of files from EMBL entries, 'rptmgenerator' which generated ".rptm" files from RepeatMasker output files, and 'eponinehits' which converted output from the Eponine promoter prediction program into ".misf" files. A number of other programs were also developed to aid in the maintenance of these files: 'reverseclone' alters co-ordinates so the clone can be plotted in the reverse orientation, whilst 'addclone' allows the three files from each clones to be added to the files from another clone. The advantage of storing information in this format was that information could be quickly replotted with additional details or to a different scale. The program 'plotclone' was developed to take input from these three files and produce a postscript plot of features listed in these files. It also allows an optional GC line to be plotted under the other information. (Appendix 4 contains more information about these programs).

2.21. The nomenclature of olfactory receptor genes and the 'ROLF' database

In recent years, a number of nomenclatures for olfactory receptor genes have been published (Zhao and Firestein, 1999, Glusman *et al.*, 2000, Zozulya *et al.*, 2001). These nomenclatures all have different advantages and disadvantages, as does the nomenclature that was used throughout this project. All of the genes described in this project follow a private system of nomenclature that was devised at the start of the project. This nomenclature (for example, hs6M1-6*01) consists of a two letter abbreviation of the species (hs, *Homo sapiens*), followed by the number giving the chromosomal location (in this case chromosome 6, although "U" was used where the location is unknown). This is followed by a letter and number indicating whether the gene shares amino acid identity with the olfactory receptors from the MOE (M1), or with one of the three types of pheromone receptor (V1, V2 or V3). This description is followed by a dash and an arbitrary gene identity number (here, -6). Finally, where the gene has alleles, an asterix and an additional number (*01) describes the corresponding allele, and if the gene is a pseudogene (P), fragment (F), or pseudogene fragment (PF) this is indicated by the appropriate letter. The database created as part of this project utilizes this new nomenclature but links to the public databases are maintained via accession numbers and the position of the gene within the accession number, where applicable. Links to other olfactory receptor gene databases were not built in, owing to time constraints, but where a gene has been extensively studied it has been referred to by the name given in the original literature.

The database of olfactory receptor genes was initially set-up through a search of the EMBL nucleotide database using 'SRS' (Sequence Retrieval Server) and searching for key words. A BLAST-searchable database was created using either the 'gcgtoblast' program (from the 'GCG' package of computer programs (Womble, 2000)) or the 'formatdb' program (part of the 'blastall' suite of programs). A small script ‘Rolfp’ was used to access the database, producing a results file in the directory from which the search was made.

2.22. Gene-finding approach for olfactory receptor genes

Analysis of DNA through running the sequence through the ‘AceDB’ system or the ‘NIX’ suite of programs is one way of identifying genes, but in the large scale analysis of olfactory receptor genes within the human genome, this approach for all clones considered to contain an OR gene would have been impractical. The method adopted for identifying OR genes was therefore to take clones that had been identified as containing olfactory receptor-like sequences (screened using ‘BLAST’) and using the dot-matrix program ‘Dotter’ to identify regions that were positive for OR genes. In the case of OR genes, regions of approximately 2 Kb were identified as positive where a row of high scores ran diagonally across a region of sequence.

In order to extract OR genes from regions of sequence, the ‘olfgrab’ program was developed. This program performs a number of steps: (i) extracts the region of sequence from the relevant clone, (ii) creates 6 files containing each translation of this piece of sequence, (iii) takes the 6 protein files and uses BLAST to compare these files against the database of olfactory receptor genes, (iv) displays the scores from these 6 BLAST searches and (v) creates a file containing the DNA and the highest scoring protein translation. (Appendix 2). This translation and DNA sequence was then manually edited to show only the olfactory receptor gene, and the program ‘olfproducer’ was used to produce a file containing the protein and the nucleotide sequence from this file.

2.23. Analysis of olfactory receptor protein structure

Putative protein features were identified using the 'PIX' suite of programs at the *HGMP*. 'PIX' is similar to 'NIX' in that it runs the protein sequence through a number of programs and outputs all the results from these programs in a graphical form. 'PIX' uses programs searching other protein databases (*'Pfam', 'BLOCKS', 'PRINTS', 'PROSITE'*), predicting protein sorting signals (*'Psort'*), predicting protein secondary structure (*'DSC', 'Simpa96', 'Phd', 'Predator'*) and predicting coiled coils regions (*'COILS'*). In addition it also uses programs predicting transmembrane domains (*'Tmpred', 'Tmap', 'DAS'*), helix-turn-helix motifs ('HTH'), cleavage sites (*'Signal', 'Sigcleave'*), antigenic sites (*'Antigenic'*) and proteolytic enzyme sites (*'Digest'*). Owing to the lack of data about proteins, 'PIX' is less useful in annotating proteins than 'NIX' is in annotating DNA sequence. It did, however, provide estimates for the placement of transmembrane domains within the olfactory receptor proteins. For the consensus olfactory receptor sequence, the placement of transmembrane domains was calculated according to the placement of these domains in each individual protein. This method for predicting transmembrane domains is subject to a number of limitations, notably that these type of prediction programs have a range of 26% to 69% in accurately predicting all transmembrane domains within a protein (Moller *et al.*, 2001). The predictions that were made for OR proteins, however, can be considered to be generally correct since an evaluation of protein prediction programs revealed that one of the prediction programs used in the analysis accurately predicted transmembrane domains in 85% of G-protein coupled receptors tested (Moller *et al.*, 2001).

2.24. Other programs/scripts used throughout the thesis

In addition to the scripts and programs already described, a number of other scripts and programs were used on a regular basis. From EMBOSS suite of programs (Rice *et al.*, 2000), a number of programs were used, including ‘seqret’, a reformatting program; ‘revseq’, which complements a sequence; ‘transeq’, which translates a specific frame and ‘est2genome’ which aligns a set of spliced nucleotide sequences to an unspliced genomic DNA sequence were all used regularly. ‘Water’, a program that uses a modified Smith-Waterman algorithm to produce a local alignment between 2 sequences (DNA or protein) was also used frequently. In addition, I developed a number of scripts, listed in Appendix 4. (This also includes 5 scripts developed by Roger Horton (Chromosome 6 database curator, The Sanger Centre)).

2.25. Genome browsers

The dramatic increase in the amount of sequence data in the public domain led to a number of new bioinformatic resources being developed to provide places to store and annotate this data. Two of these resources were used in this project to provide SNP data for the human MHC extended class I region, and mouse sequence from OR clusters. *‘Ensembl’* is a joint project between EMBL-EBI and the Sanger Institute (Hubbard *et al.*, 2002) and the *UCSC human genome project* is based at UC Santa Cruz (Kent et al, unpublished, Kent and Haussler, 2001).

2.26. Analysis of regulatory regions within OR clusters

2.26.1 ‘Promoter Inspector’, ‘Eponine’ and ‘Transfac’

‘Promoter Inspector’ (Scherf *et al.*, 2000) is an algorithm that relies on the assumption that promoters are embedded into a common genomic context. It assumes that this context can be detected by classifying varying oligonucleotide sequences as ‘promoter’ or ‘non-promoter’-like. The classification was made possible by an original training period in which the algorithm was given access to eukaryotic promoter sequences (from the eukaryotic promoter database (EPD) V

60.0 (Cavin Perier *et al.*, 1998)) and vertebrate exon and intron sequences (randomly extracted from GenBank). 'Promoter Inspector' is currently considered to be 43% accurate in its prediction of promoter sites, and has a dramatically lower rate of false positives compared to older promoter prediction programs. 'Eponine' (Down and Hubbard, 2002) also relies on the detection of a common genomic environment. Classification models that distinguish between promoter and non-promoter-like sequence were trained on mammalian promoters from the EPD (positive sequences) and on an equal number of random sequence fragments from human chromosome 20 that were not annotated as promoters. 'Eponine' has been predicted to have a detection sensitivity of 40%. The false positive rate is also comparable to the rate estimated for 'Promoter Inspector': generally about 45-55% of hits can be considered false positives.

'TRANSFAC' is a comprehensive database containing transcription factors, their genomic binding sites and DNA-binding profiles (Wingender *et al.*, 2000). It can be accessed using programs such as '*MatInspector'* (Quandt *et al.*, 1995), which works with binding profiles generated from the 'TRANSFAC' matrix. 'MatInspector' generates two scores for a sequence matching to a transcription factor, a value for similarity to the matrix and a value for similarity to the core (which is made up of the four consecutive bases within the matrix which show the highest amount of conservation).

2.26.2. DNA Block Aligner ('DBA')

'DBA' (Jareborg *et al.*, 1999) is a program that aligns 2 sequences assuming that sequences share blocks of conservation separated by large and varied lengths of DNA within the 2 sequences. The conserved blocks may have 1 or 2 gaps in them and the amount of conservation can be specified (ranging from 65% upwards). This approach was designed for comparing the upstream regions of genes both between and within species: conserved blocks may be important in regulating the gene.

2.27. Analysis of comparative regions in the human and mouse OR clusters: 'PipMaker'

'*PipMake*r' (Schwartz *et al.*, 2000) is a program that is able to align two long DNA sequences to identify conserved sequences. These conserved sequences are shown in a graphical form as a percentage identity plot (PIP). PIPs are plotted with one sequence along the horizontal axis: this sequence can be labelled with features such as exons and repeats. Sequences conserved within the horizontal sequence compared to the other input sequence are indicated by a vertical line positioned at the appropriate place along the horizontal axis. The extent of conservation is indicated by the position of the vertical line along the vertical axis: conservation ranges from 50% (bottom of vertical axis) to 100% (top of vertical axis). Files containing information about these conserved sequences in textual form are also produced by 'PipMaker,' along with the corresponding dot-matrix plot of the region (similar to plots produced by the 'Dotter' program).

2.28. Phylogenetic analysis

2.28.1. Protein alignments

All protein alignments were performed using the '*ClustalW*' program (Thompson *et al.*, 1994). This is a progressive multiple sequence alignment method that 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. It also varies amino acid substitution matrices at different alignment stages according to the divergence of the sequences to be aligned, and 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. In the case of most olfactory receptor alignments, the high number of closely related sequences means that alignments produced by this program tend to be very close to ideal, although all alignments were examined in the program 'belvu' (Sonnhammer, unpublished) before they were used. Alterations to protein alignments were made using the program 'jalview,' (Clamp, unpublished) or alterations were made in Excel after converting a alignment FASTA file into a tabbed alignment using the program 'ProAlnExcel' (Appendix 4).

2.28.2. Phylogenetic tree production

With the exception of the large tree of 716 OR proteins, all phylogenetic trees were constructed using the program 'phylo_win' (Galtier *et al.*, 1996). This is a graphical interface for molecular phylogenetic inference, which can perform neighbor-joining and parsimony methods.

2.28.2.i. The neighbor-joining method and distance calculation methods.

The neighbor-joining (NJ) method (Saitou and Nei, 1987) is a distance based method which uses an algorithm to convert pairwise distances between sequences into a matrix, from which branching order and branch lengths are computed. Advantages of this method include that it is a relatively computationally light process and that only one tree is produced. The disadvantages of this approach stem from these two advantages: producing only one tree means other trees are not evaluated, and the algorithm may not provide an accurate depiction of historical events. Methods to calculate distances for the neighbor-joining method also have a number of problems associated with them. Distance calculation methods available for use in the 'phylo_win' package include the observed divergence, which is simply the observed percent of differences between the compared sequences, the Poisson Correction which attempts to corrects for multiple substitutions according to a one-parameter model, and the PAM distance which corrects for multiple substitutions according to Dayhoff's PAM matrix (Dayhoff, 1976).

In the trees created in this project, the PAM matrix was used as this provides the most complex model of protein evolution, in providing a measure of probability calculating how likely the amino acid in one sequence is likely to change in the amino acid in the other sequence. These probabilities were based on a subset of closely related proteins that were organized 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. This matrix is based on a number of assumptions which will all cause problems when applying this matrix in creating a phylogenetic tree: (i) each amino acid site is equally mutable (ii) the frequency of amino acid changes that would require two nucleotide changes is higher than would be expected by chance (iii) the matrix is based on a small set of closely related proteins (there are other updated matrices based on more protein sequences that could be used (Gonnet *et al.*, 1992, Jones *et al.*, 1992)). The advantage of the PAM matrix, however, is that the frequency of changes was averaged across conserved and non-conserved blocks within the protein (Henikoff and Henikoff, 1992). In view of the fact that olfactory receptor proteins are made up of conserved and non-conserved blocks of protein sequence, therefore, it seemed appropriate to use a matrix based on both conserved and non-conserved region of sequence. Ideally, an olfactory receptor-specific matrix should be used to generate phylogenetic trees but time constraints did not allow for the development of this matrix. In ‘phylo_win’, the PAM matrix is applied using the algorithm of the program ‘PROTDIST’ (Felsenstein 1993, unpublished) of the ‘PHYLIP’ package (Felsenstein 1989).

2.28.2.ii. The (maximum) parsimony method

The second method used to generate phylogenetic trees was parsimony, which is a method that uses the multiple alignment directly rather than using an algorithm based on a calculated distance. Parsimony (Fitch, 1971) is based on the assumption that the most likely tree is the tree that requires the smallest number of changes to explain the data in the alignment. This assumes all the data in the alignment share a common evolutionary origin, and the smallest number of evolutionary steps required to produce the data have been performed. The method calculates a number of trees that could be produced from the data and chooses the tree requiring the lowest possible number of changes to explain the data. In an ideal type situation all possible trees would be calculated and evaluated, however, in reality it would too computationally intensive to generate all possible trees and so parsimony methods generally rely on a heuristic strategy, in which an initial tree is selected and rearrangements are then made to this tree to find the most parsimonious tree. Maximum parsimony has a number of advantages: it does not reduce sequence information to a single number, it tries to provide information on the ancestral sequences and it evaluates different trees. Disadvantages to maximum parsimony include that it is slow in comparison with distance methods, it does not use all the sequence information (only informative sites are used), and it does not correct for multiple mutations that may have occurred in the evolutionary process. Parsimony can also be biased in dealing with among-site variation, it can group sequences with a higher number of changes together rather than considering relatedness, and it can generate a number of trees that are equally parsimonious. Whenever used, these limitations should be considered especially with distantly related protein sequences or sequences of a different function. In the case of olfactory receptor genes, all protein sequences are considered to be derived from a common ancestor and it is assumed that these sequences possess a similar function. The assumed equality of mutational and selectional pressures on olfactory receptor genes means that some of the limitations of this method are less acute when performing reconstructions of this family's evolutionary history. In 'phylo_win' the 'PROTPARS' program from the 'PHYLIP' package is used. (Felsenstein 1989).

2.28.2.iii. Bootstrapping phylogenetic trees

Boot-strapping (Felsenstein, 1985) is a method available in 'phylo_win' that allows the reliability of groupings within a tree to be evaluated. This method 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. Bootstrap values of over 70% were considered to represent reliable groupings, whilst groupings defined with low bootstrap values at their branch points were generally disregarded, although low bootstrap values at interior branches do not necessarily mean the entire phylogenetic tree is worthless. Every phylogenetic tree is the best tree obtainable using a specific method, and computer simulations have shown that that branching patterns of an inferred tree may be correct even if they are not supported by high bootstrap values (Nei and Kumar, 2000).