*Published online 30 May 2008* *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W149–W153***
*doi:10.1093/nar/gkn293*

# Uprobe 2008: an online resource for universal overgo hybridization-based probe retrieval and design†

**Robert T. Sullivan[1], Caroline B. Morehouse[1], NISC Comparative Sequencing Program[2] and James W. Thomas[1,*]**

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322 and
[2]Genome Technology Branch and National Institutes of Health Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

Received January 23, 2008; Revised April 17, 2008; Accepted April 29, 2008

## ABSTRACT

**Cross-species sequence comparisons are a prominent method for analyzing genomic DNA and an ever increasing number of species are being selected for whole-genome sequencing. Targeted comparative genomic sequencing is a complementary approach to whole-genome shotgun sequencing and can produce high-quality sequence assemblies of orthologous chromosomal regions of interest from multiple species. Genomic libraries necessary to support targeted mapping and sequencing projects are available for more than 90 vertebrates. An essential step for utilizing these and other genomic libraries for targeted mapping and sequencing is the development of the hybridization-based probes, which are necessary to screen a genomic library of interest. The Uprobe website (http://uprobe.genetics.emory.edu) provides a public online resource for identifying or designing 'universal' overgo-hybridization probes from conserved sequences that can be used to efficiently screen one or more genomic libraries from a designated group of species. Currently, Uprobe provides the ability to search or design probes for use in broad groups of species, including mammals and reptiles, as well as more specific clades, including marsupials, carnivores, rodents and nonhuman primates. In addition, Uprobe has the capability to design custom probes from multiple-species sequence alignments provided by the user, thus providing a general tool for targeted comparative physical mapping.**

## INTRODUCTION

Genomic resources which were once restricted to humans and traditional model systems, such as whole-genome sequence assemblies and large-insert genomic libraries, are now being developed for an ever expanding list of species. Driven in part by the demonstrated utility of interspecies sequence comparisons for genome annotation and decreasing costs of DNA sequencing, these new genomic tools are providing an ever expanding platform for studies in comparative genomics. While whole-genome shotgun sequencing has largely replaced the clone-by-clone-based method to sequence the human genome (1,2), high-quality targeted genomic sequencing of chromosomal regions of interest derived from clone-based physical maps provides an alternative approach for comparative genomic sequencing. In particular, large-insert bacterial artificial chromosome libraries (BAC libraries) that are currently required to undertake targeted mapping and sequencing projects are available for many diverse species, including more than 90 vertebrate (for example, see http://bacpac.chori.org). Importantly, a substantial fraction of genomic libraries represent genomes that have not been subjected to whole-genome sequencing. Targeted physical mapping and sequencing can therefore be used to augment existing comparative genomic data sets from chromosomal regions of interest (3,4) and generate high-quality sequence data in regions of incomplete or low-quality assemblies generated by whole-genome shotgun sequencing (5). The creation of physical maps also provides direct access to sequenced BAC clones that can be used in a variety of experimental settings. It is therefore important that efficient methods be available for undertaking targeted comparative mapping and sequencing projects.

*To whom correspondence should be addressed. Tel: +404 727 9751; Fax: +404 727 3949; Email: jthomas@genetics.emory.edu
†The list of GenBank accession numbers are given in the Appendix.

Targeted physical maps are typically constructed by screening a BAC library with a set of hybridization-based probes that correspond to the chromosomal region of interest. Species-specific 'overgo' probes, which are 36–40 bp radioactively labeled double-stranded DNA probes (6), have been shown to be exceptionally useful for screening genomic libraries (7). Overgo design programs include, Overgo Maker, (http://genome.wustl.edu/tools/software/overgo.cgi), which takes as input non-aligned *.fasta formatted files, and OligoSpawn (http://138.23.191.145/) (8), which can design either 'unique' or 'popular' overgo probes that respectively occur once, or more than once in a provided set of sequences. However, for a species for which a BAC library is available but for which no large-scale DNA sequence resources are available, the design of species-specific overgo probes is often not feasible. Previously, we showed that by considering sequence conservation in the design of overgo probes, 'universal' probes could be developed for efficiently screening one or more genomic library (9). Universal overgo probes therefore alleviate the need for species-specific probes and can be used in the construction of physical maps of orthologous regions from multiple species in parallel (9). To make this method more accessible, we systematically applied the universal probe design concept to whole-genome alignments and have disseminated the resulting probe sets for screening mammalian and bird/reptile BAC libraries through a website called Uprobe (http://uprobe.genetics.emory.edu) (10).

Here, we report the creation of three new predesigned whole genome sets of universal probes for screening marsupial, rodent and carnivore genomic libraries, advanced capabilities for searching the predesigned universal probe sets, a tool for the on-demand design of universal overgo probes for screening ape and old world monkey genomic libraries, and finally the capability for the custom design of universal overgo probes from user supplied DNA sequence alignments.

## OVERVIEW OF THE UNIVERSAL OVERGO DESIGN PROCESS

An overview of the steps taken in the design of universal overgo probes by the Uprobe website is illustrated in Figure 1. Briefly, an entire genome, or target region within a genome, is selected for probe design. Pairwise or multiple-species whole-genome alignments are processed with the nsoop_v2 algorithm, which incorporates the phylogenetic relationship among the species to identify the most conserved 36-mers suitable for overgo probes that can be used at a standard hybridization and washing temperature of 58°C (i.e. 36-mers with a GC content between 44% and 56%) (9,10). Potential overgo probes are then compared back to the genome of origin and/or additional genomes to identify probes which will likely hybridize to a single locus in the genome (i.e. are unique), or more than one location (i.e. are nonunique). A subset of the unique probes that are optimally spaced along the chromosome for constructing a probe-content physical map of the chromosomal region of interest is then selected, and the primer sequences and accompanying information for each overgo probe is provided to the user.

## PREDESIGNED WHOLE-GENOME PROBE SETS FOR MARSUPIALS, RODENTS AND CARNIVORES

To expand on the predesigned whole-genome sets of universal probes for screening mammalian and bird/reptile genomic libraries available at the Uprobe website, three additional subclades of mammals, marsupials, carnivores and rodents, were selected for the development of whole-genome universal probe sets. In each case, whole-genome alignments were downloaded from the UCSC Genome Browser (11) and used as the basis for universal probe design. The marsupial universal probe set was developed from the assembled opossum genome (monDom1) in regions conserved with the human genome (hg17), whereas the rodent and carnivore probe sets were derived from the mouse (mm5) and dog (canFam1) genomes, respectively, in regions conserved between the mouse, rat (rn3), human and dog genomes. Unique probes from the marsupial, rodent and carnivore probe sets are spaced on average every 29, 9 and 6 kb, respectively (Table 1) (10), and with the exception of the marsupial probe set that had the lowest density of probes, provide coverage of the genome comparable to that of the previously described mammalian and bird/reptile probe sets (Table 1). Note that like the previously described mammalian and bird/reptile whole-genome probe sets, the coverage across the genome for the new probe sets are nonuniform and dependent on a variety of factors including the density of conserved sequences (data not shown) (10). To validate the design properties used to develop each probe set, samples of 48 representative probes consisting of 4–8 linked probes from 10 different regions of the genome were experimentally tested by screening one or more genomic libraries from the target clade of interest. The resulting probe success rates, as defined by the percentage of probes that successfully identified at least two but fewer than 20 clones in each ~10× library, were as follows: marsupial—75% for wallaby (ME_KBa); rodent—31% for squirrel (VMRC-20) and 83% for deer mouse (CHORI-233); carnivore—73% for clouded leopard (CHORI-87) (Table 1). To assess the combined ability of the clusters of linked probes to specifically isolate the orthologous genomic regions of interest in each species, probe-content and restriction-enzyme fingerprint mapping data were used to select representative sets of BAC clones for sequencing (Table 1). Two-thirds of the sequenced marsupial BAC clones and 100% of rodent and carnivore sequenced BAC clones mapped to the targeted orthologous region. The genome coverage, probe success rates and sequence validation therefore demonstrate that the marsupial, rodent and carnivore whole-genome predesigned universal probe sets can be successfully used for targeted physical mapping in each of these clades. Note that more details regarding the probe design process and experimental validation of each of these probe sets, including files describing the specific probes that were

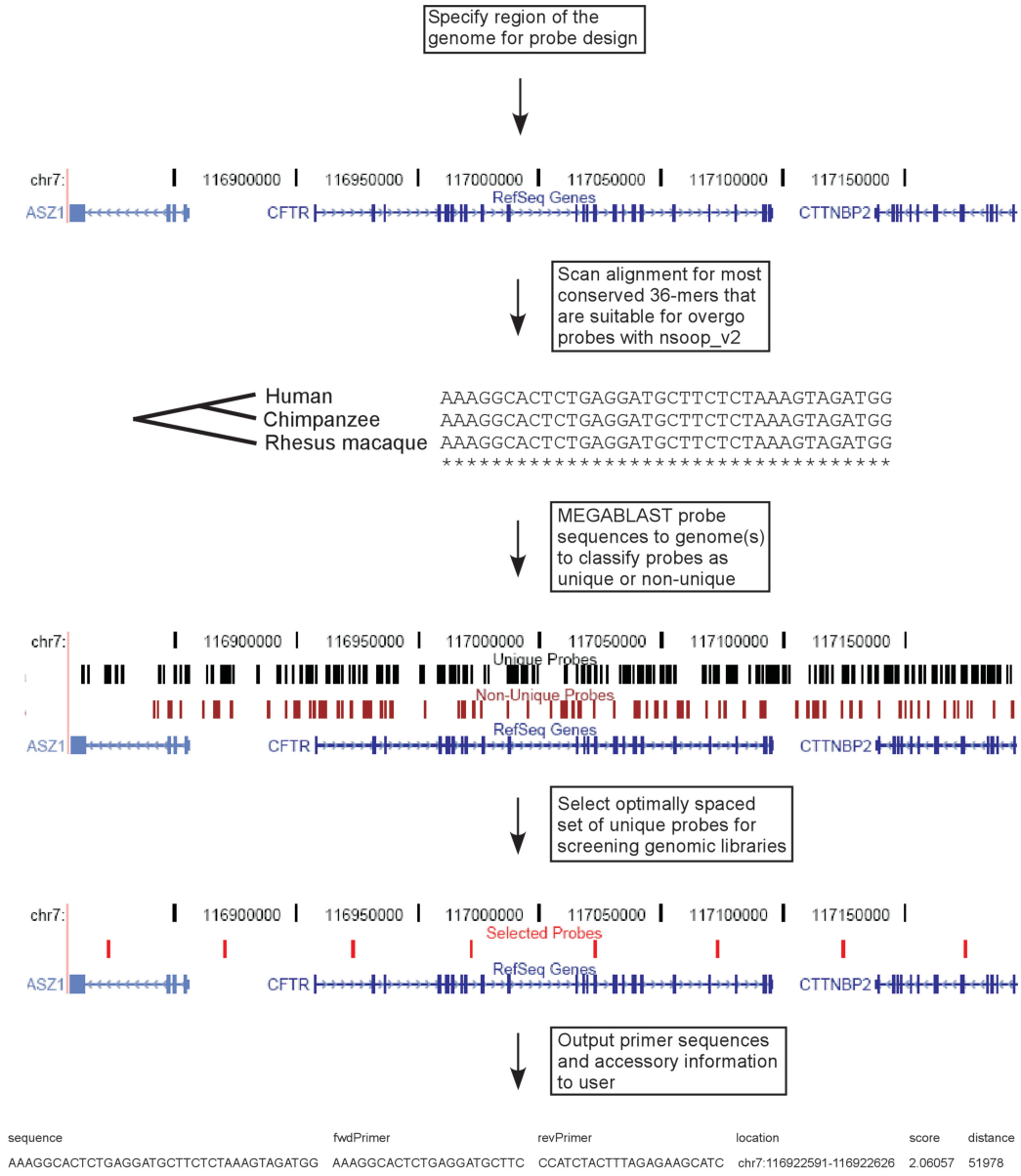


**Figure 1.** Overview of the universal overgo probe design process.

tested and the mapping results, is provided on the Uprobe website (http://uprobe.genetics.emory.edu/direction.php).

## BATCH AND CROSS-SPECIES OPTIONS FOR UNIVERSAL PROBE RETRIEVAL

The Uprobe website was originally designed to support a variety of text-based queries, such as chromosome location, gene name or GenBank accession number, to identify the genomic region of interest and the corresponding universal probes that map to that interval of the genome. However, batch queries were not supported, and queries needed to be in reference to the genome from which the probes were designed. To expand the ability of users to search for universal probes of interest, we implemented a pair of advanced search options (http://uprobe.genetics.emory.edu/advanced_search.php) that specifically are designed to handle batch queries or cross-species searches. With the batch query tool, probes from multiple loci can be searched for simultaneously by inputting a list of text queries, thus eliminating the need to sequentially enter queries for single loci. The cross-species tool provides the capability of directing a query not only at the genome from which the probes were derived, but also one of the comparative genomes used in the probe design process. For example, a user may want to initiate a comparative mapping and sequencing effort in marsupials orthologous to a defined chromosomal region in the human genome. With the cross-species application it is possible to retrieve probes from the marsupial probe set based on their syntenic location in the human genome. In other words, the cross-species application will translate a user-defined interval in the human genome to the

orthologous position in the marsupial genome, thereby identifying probes from the chromosomal region of interest. These enhanced query capabilities, which were in part in response to feedback from the public, provide useful additions to the search options available to the Uprobe website.

## ON-DEMAND DESIGN OF UNIVERSAL PROBES FOR APES AND OLD WORLD MONKEYS

BAC libraries representing 16 different species of nonhuman primates are now available (http://bacpac.chori.org) (12), and represent a unique and valuable resource for reconstructing the history of the human genome, as well as for comparative studies of the sequenced genomes of nonhuman primates used in biomedical research, like Rhesus macaque (old world monkey) (13), and marmoset (new world monkey). In order to facilitate targeted mapping and sequencing efforts in nonhuman primates, we developed an on-demand universal probe design tool focused on these species. Similar to the query interface to the predesigned universal probe sets, text-based queries of chromosome location, gene name, accession number or keyword are used to initiate the probe selection process. Universal probes are then designed with a set conservation criteria optimized for the clade of interest and returned to the user via an email. However, unlike the predesigned probe set application, the on-demand tool designs probes on-the-fly in response to each query. A major benefit of the on-demand probe design pipeline is that it formed the basis of a more flexible probe design tool that could be implemented in other settings (i.e., the custom probe design tool, which is described in the next section).

At the current time, the on-demand tool supports the design of universal probes for screening ape and old world monkey genomic libraries. The universal probes for this set of species are derived from the human genome (hg18) conserved sequences identified by using the human, chimpanzee (panTro1) and macaque (rheMac2) whole-genome alignment downloaded from the UCSC Genome Browser. Similar to the cross-species advanced search option described earlier, queries using this tool can be directed at locations in either the human or macaque genome. The estimated average probe spacing for ape and old world monkey universal probes is 1 unique probe every 2 kb, and the experimentally determined success rates for the probe design criteria were: 94% gibbon (CHORI-271), 94% Japanese macaque (CHORI-270), 83% baboon (RPCI-41) and 83% colobus monkey (CHORI-272) (Table 1). Compared to the predesigned universal probe sets, the ape and old world monkey probes are found at a greater frequency in the genome and have a higher minimum probe success rate (Table 1). Both these observations are a consequence of the relatively low divergence among apes and old world monkeys compared to the more evolutionary diverse groups of species targeted previously. In addition, sequencing of representative BAC clones indicated that the combined specificity of the probes was very high and ranged from 86% in the colobus monkey to 100% in gibbon and the other two old world monkeys (Table 1). Thus, this application can provide a highly efficient method of designing probes for targeted physical mapping and sequencing in apes and old world monkeys. Future upgrades to the Uprobe site will include expanding the on-demand tool to other clades of nonhuman primates.

**Table 1.** Available probe sets and tools for universal probe retrieval and design

| Species group | Average genome-wide probe spacing (probe/kb) | Probe success rates (%)[a] | Aggregate probe specificity (%)[b] |
|---|---|---|---|
| Birds and reptiles[c] | 1/14 | 41–98 | 90–100 |
| Mammals[c] | 1/7 | 36–95 | 66–100 |
| Marsupials | 1/29 | 75 | 66 |
| Rodents | 1/9 | 31–83 | 100 |
| Carnivores | 1/6 | 73 | 100 |
| Apes and old world monkeys | 1/2[d] | 83–94 | 86–100 |
| Custom | NA[e] | NA[e] | NA[e] |

See http://uprobe.genetics.emory.edu/direction.php for more details, including summary files of the test probes and mapping results.
[a]Probe success rates were derived from experimental testing of sets of 48 probes on the target clade. Note that the criteria used to estimate the probe success rates was slightly different than described in the text for the previously published probe sets, mammals and birds and reptiles (10).
[b]The percentage of sequenced BAC clones that mapped to the targeted orthologous region was determined based on previously published criteria (9,10). The number of clones sequenced for each species is: wallaby ($n = 3$), squirrel ($n = 6$), deer mouse ($n = 18$), clouded leopard ($n = 16$), gibbon ($n = 8$), Japanese macaque ($n = 7$), baboon ($n = 5$) and colobus monkey ($n = 7$).
[c]Previously published probe sets (10).
[d]Estimated from a sample of $n = 51\,250$ kb regions that are representative of the genome-wide distribution of divergence between human and chimpanzee.
[e]The custom probe tool designs probes from user provided sequence alignments.

## CUSTOM UNIVERSAL PROBE DESIGN TOOL

A limitation of our predesigned probe sets and on-demand probe design tool is that they are not capable of continuously incorporating the ever expanding list of genome sequence assemblies (14) in the probe design process. In addition, while our probe design efforts have been focused on a select set of vertebrates, the universal probe design concept could be applied to any set of species for which comparative sequencing data are available. We therefore developed a custom universal probe design tool based on our validated on-demand probe design pipeline that allows the custom design of universal probes based on user-provided sequence alignments. Initiated by the upload of an alignment file to the Uprobe website, the custom universal probe design tool walks the user through the necessary steps of probe design illustrated in Figure 1. Help and example files are provided for each step of the process on the website. Though primarily designed to process maf formatted alignment files that can be directly downloaded for any given region of a reference genome from the UCSC Genome Browser, the custom probe design tool will also accommodate probe design from

fasta, blastz, clustal, phylip and lagan formatted files. After going through the full set of required steps, including selection of a reference sequence for probe design and comparative species, confirmation of a phylogenetic tree describing the set of species to be used in probe design, selection of optimal probe spacing and type of probe (unique or nonunique), the best set of conserved probes is returned to the user via email. Note, that unlike the predesigned probe sets and probes generated from the on-demand tool, no minimum conservation threshold is implemented in the custom pipeline. However, an output file is generated that includes the alignment information and conservation 'score' for each overgo that can be postprocessed by the user to identify sequences conserved at a given threshold. The custom probe design tool therefore allows the public to apply our proven universal overgo probe design pipeline for the development of universal probes from DNA sequence alignments between any given set of species.

## CONCLUSION

The Uprobe website, first described in 2005 (10) has served as a unique online resource for universal overgo-hybridization retrieval and design. Since the initial development of Uprobe, many new genomes have been sequenced, thus greatly expanding the capability to design universal probes. Here, we have described the updates to the Uprobe site, including new search options, as well as capabilities that have exploited these new sequence resources, i.e. new whole-genome probe sets for screening marsupial, rodent and carnivore genomic libraries, an on-demand tool for universal probe design for screening nonhuman primate genomic libraries, and a custom probe design tool. As such, Uprobe provides a novel online tool for targeted comparative physical mapping in vertebrates and other species.

## ACKNOWLEDGEMENTS

The authors wish to thank the NISC Comparative Sequencing Program for their testing of the website and the specific contributions of E. D. Green, R. T. Blakesley, G. G. Bouffard and P. J. Thomas, and G. K. Tharp for his help debugging the website. This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. RTS, CBM and JWT were supported by a grant from the National Institutes of Health (R24RR022239). Funding to pay the Open Access publication charges for this article was provided by a grant from the National Center for Research Resources (R24RR022239).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
3. Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M., Margulies,E.H., Blanchette,M., Siepel,A.C., Thomas,P.J., McDowell,J.C. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
4. ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
5. Gordon,L., Yang,S., Tran-Gyamfi,M., Baggott,D., Christensen,M., Hamilton,A., Crooijmans,R., Groenen,M., Lucas,S., Ovcharenko,I. *et al.* (2007) Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Res.*, **17**, 1603–1613.
6. Vollrath,D. (1999) In Birren,B., Green,E.D., Hieter,P., Klapholz,S., Myers,R.M., Riethman,H. and Roskams,J. (eds), *Genome Analysis: A Laboratory Manual, Volume 4: Mapping Genomes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 187–215.
7. McPherson,J.D., Marra,M., Hillier,L., Waterston,R.H., Chinwalla,A., Wallis,J., Sekhon,M., Wylie,K., Mardis,E.R., Wilson,R.K. *et al.* (2001) A physical map of the human genome. *Nature*, **409**, 934–941.
8. Zheng,J., Svensson,J.T., Madishetty,K., Close,T.J., Jiang,T. and Lonardi,S. (2006) OligoSpawn: a software tool for the design of overgo probes from large unigene datasets. *BMC Bioinform.*, **7**, 7.
9. Thomas,J.W., Prasad,A.B., Summers,T.J., Lee-Lin,S.Q., Maduro,V.V., Idol,J.R., Ryan,J.F., Thomas,P.J., McDowell,J.C. and Green,E.D. (2002) Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.*, **12**, 1277–1285.
10. Kellner,W.A., Sullivan,R.T., Carlson,B.H. and Thomas,J.W. (2005) Uprobe: a genome-wide universal probe resource for comparative physical mapping in vertebrates. *Genome Res.*, **15**, 166–173.
11. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
12. Eichler,E.E. and DeJong,P.J. (2002) Biomedical applications and studies of molecular evolution: a proposal for a primate genomic library resource. *Genome Res.*, **12**, 673–678.
13. Gibbs,R.A., Rogers,J., Katze,M.G., Bumgarner,R., Weinstock,G.M., Mardis,E.R., Remington,K.A., Strausberg,R.L., Venter,J.C., Wilson,R.K. *et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.
14. Miller,W., Rosenbloom,K., Hardison,R.C., Hou,M., Taylor,J., Raney,B., Burhans,R., King,D.C., Baertsch,R., Blankenberg,D. *et al.* (2007) 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, **17**, 1797–1808.

## APPENDIX

The GenBank accession numbers are as follows: AC217456.1, AC217457.1, AC217464.1-AC217469.1, AC217471.1, AC217474.1-AC217478.1, AC217481.1, AC217485.1, AC217592.1-AC217594.1, AC217599.1, AC217600.1, AC217614.1, AC217621.1, AC217738.1-AC217740.1, AC217745.1-AC217751.1, AC217765.1, AC217766.1, AC217856.1, AC217857.1, AC217864.1, AC217879.1, AC218054.1-AC218056.1, AC218064.1, AC218065.1, AC218076.1-AC218078.1, AC218087.1, AC218890.1, AC218891.1, AC222517.1, AC222544.1, AC222548.1, AC222549.1, AC225042.1, AC225044.1, AC225053.1, AC225055.1, AC225056.1, AC225057.1, AC217748.1, AC225272.1-AC225275.1, AC225257.1-AC225259.1, AC225265.1, AC225266.1, AC225276.1, AC225304.1.