Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing

Supplementary text 2 - Classification of gene loci

Missing Loci

There are 18 gene loci not present in the genome sequence meaning that no portion of the mRNA sequences from that loci could be aligned at 98% base pair identity. A full list of these 18 loci are provided in Supplemental Table 6.

For each of these missing loci, we attempted determine its location using the locations of orthologous gene loci in the mouse and rat genomes. We first attempted to align each missing mRNA sequence to the latest assemblies of the mouse and rat sequence and were successful in 10 instances. Using BLASTZ alignments (Schwartz et al., 2003) between the human and rodent genomes, further refined by the chaining and netting algorithm available in the UCSC Genome Browser (Kent et al., 2002), we are able to map these missing loci to known gaps in the human genome sequence. Two additional loci can be reasonably associated with known gaps. MMP23A is located in a duplicated region on chromosome 1 that includes a second matrix metalloproteinase gene, MMP23B. It is, therefore, reasonable to assume that MMP23A falls into the gap that immediately precedes MMP23B. CSF2RA is known to reside in a pseudo-autosomal region in chromosomes X (Xp22.32) and Y (Yp11.3). This locus has been mapped to gap between accessioned clones AL672277 and BX296563.

Of the remaining six loci, four (DUX1, DUX2, DUX3, DUX5) are part of the double homeobox family that are homologous to 3.3 kb dispersed repeated elements. These have been mapped to acrocentric chromosomes and are possibly located on the short arm in the heterochromatic regions adja-

cent to the ribosomal DNA gene clusters (Ding et al., 1998). Another locus, ANAPC7, actually is present in clone AC144548 that is partially included in the sequence of chromosome 12. Due to a technical error, the portion of AC144548 containing ANAPC7 is not present in the final chromosome 12 sequence. The last locus is based on the mRNA sequence BC034024 that reportedly encodes a hypothetical protein. No mapping information exists for this sequence. It aligns at nearly 98% base pair identity to chromosome 1, bases 141936765-141937918. This locus, though, is most certainly a processed pseudogene. BC034024 aligns in a single block, and a poly-A tract immediately follows this alignment, both strong indications of a pseudogene. It is reasonable to assume that a functioning copy of this gene exists somewhere in the human genome, but it is not present in the current genome sequence and its location is unknown.

Partially found loci

For 81 loci represented by 127 sequences, some or all of the sequence could be aligned with at least 98% base pair identity, but no single alignment covered at least 95% of the sequence. For 46 of these loci, the lack of an alignment is directly due to part of the sequence being located in a gap or in random genome sequence that will eventually fill current gaps. In 14 of these 46 instances, the full sequence is completely present in genome sequence, but some or all of it is in the random genome sequence. Supplemental Table 7 provides a listing and locations for each of these 46 loci.

For the remaining 35 loci, evidence indicates that either a deletion in the genome sequence (31 loci) or a mis-assembly (4 loci) prevents a full alignment of the corresponding mRNA sequence(s). Supplemental Table 7 also lists these 35 loci along with their genome locations. Fosmid and BAC clone end sequence alignments often provided evidence of a genome sequence problem. For misassemblies, this includes observing pairs in which the orientation of the aligned end sequences are not complementary. For misassemblies and deletions, the alignment of one end sequence but not its pair was also indicative of a problem.

Well-aligned found loci

For the remaining 16,920 loci represented by 30,631 sequences, the mRNA sequences can be aligned at greater than 98% base pair identity with at least

95% of the sequence aligning. These are called found loci. There are a few of these found loci that did have some significant ambiguities.

Recently duplicated loci with ambiguous mRNA placements

For 112 loci represented by 192 sequences, the mRNA can be aligned nearly equally well in more than one location in the genome sequence. For 111 of these 112 loci, the alignments are contained in identified regions of recent segmental duplications that are a minimum of 1000 bases in length with greater than 90% sequence identity (Bailey et al., 2002). The alignments of the last loci containing mRNA NM_014058 (DESC1), are also in a region that shows strong evidence of duplication, though probably not as recently. The alignments differ by a single base pair mismatch, and one alignment contains 4 more intronic bases. They both are followed by a UDP glycosyltransferase 2 family gene of nearly identical size and structure (UGT2B17 and UGT2B15), though these are different enough to be clearly distinct. For the sake of creating a non-redundant set of loci, a single location was chosen for each of these based on the maximizing the number of base pair matches in the annotated coding regions.

Weakly aligned mRNA sequences

There are 32 mRNA sequences that do not align with 98% identity over at least 95% of their length, but for which a strong, nearly complete alignment can be made at a locus where one or more other mRNA sequences align satisfying this criteria. Fourteen of these sequences come from an immunoglobulin or immunoglobulin-related locus. Another fifteen code for major histocompatibility genes. Both of these genes families are known to be highly polymorphic in the population, and therefore these sequences most likely represent alleles of the loci represented in the human genome sequence. Two more mRNA sequences, NM_003293 and BC028059, encode the tryptase gene TPS1. Both sequences align completely, but at less slightly than 98% base pair identity at a locus in the genome sequence represented better by mRNA BC029356. The tryptase family of genes cluster together at 16p13.3 and are 98-99% identical in amino acid sequence with the exact number of genes and alleles in this family not precisely known (Pallaoro et al., 1999). Therefore, NM_003293 and BC028059 are likely to be alleles of this locus.

NM_012313, which encodes KIR2DS3, is part of the family of killer cell immunoglobulin-like receptor genes found at 19q13. The best alignment for this sequence reports slightly less than 98% base pair identity, and is at the same position as KIR2DS5 represented by NM_014513. The whole KIR locus has been reported to be polygenic and polymorphic (Uhrberg et al., 1997), so NM_012313 could reasonably be considered an allele of the KIR2DS5 locus and not a distinct gene.

Other highly polymorphic loci

Three loci are considered found, though it is not entirely clear that they are completely and accurately represented in the genome sequence. First, NM_005577, which encodes lipoprotein, Lp(a) (LPA), is nearly 14,000 base pairs long and contains 38 copies of a 233 base pair Kringle domain. It has been shown that this locus is highly polymorphic with the number of Kringle domains varying in the population (Rosby and Berg, 2000). NM_005577 aligns at 100% base pair identity, but with less than half of the bases aligning due primarily to the absence of some Kringle domains. It is unclear whether the genome sequence represents a viable representation of this locus.

The second mRNA, NM_002457, encodes mucin 2 (MUC2) with a sequence length of 15,700 bases. Again, the base pair identity with its location in the chromosome 11 genome sequence is almost perfect, but only 8,600 bases are aligned. MUC2 is polymorphic and contains two very large repetitive regions, the first containing many copies of a 48 base repeat and the second containing possibly over 100 copies of a 69 base repeat (Toribara et al., 1991). The unaligned region of NM_002457 spans both of these regions in a single, unaligned chunk. The BAC clone representing this region has been tagged as being unsure about the sequence covering this locus, and so the genome sequence may not be an accurate representation any allele of this locus.

Finally, a member of the cytochrome P450 family (CYP2D6) represented by NM_000106 aligns at slightly less than 98% to clone AL021878 on chromosome 22. CYP2D6 is known to be highly polymorphic in the population (Sachse et al., 1997), but is also reportedly near two unprocessed psuedogenes originating from this locus. This is the only mRNA sequence for this locus in either RefSeq or MGC. Several other full and partial mRNA sequences from GenBank align at this locus, most between 97-98% base pair identity attesting to the polymorphic nature of this locus. One partial mRNA sequence, X16866, does align at nearly 100% covering 3' end of the

transcript. Therefore, the genome sequence at this locus may represent a valid allele, but not the allele represented by NM_000106. It is also possible, though, that this genomic location represents a pseudogene, and that the real locus is deleted.

References

Bailey, J., Gu, Z., Clark, R., Reinert, K., Samonte, R., Schwartz, S., Adams, M., Li, E. M. P., and Eichler, E., 2002. Recent segmental duplications in the human genome. *Science* **297**(5583):1003–1007.

Ding, H., Beckers, M. C., Plaisance, S., Marynen, P., Collen, D., and Belayew, A., 1998. Characterization of a double homeodomain protein (DUX1) encoded by a cDNA homologous to 3.3 kb dispersed repeated elements. *Hum. Mol. Genet.* **7**(11):1681–1694.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D., 2002. The human genome browser at UCSC. *Genome Res.* **12**(5):2291–2300.

Pallaoro, M., Fejzo, M. S., Shayesteh, L., Blount, J. L., and Caughey, G. H., 1999. Characterization of genes encoding known and novel human mast cell tryptases on chromosome 16p13.3. *J. Biol. Chem.* **274**(6):3355–3362.

Rosby, O. and Berg, K., 2000. LPA gene: interaction between the apolipoprotein(a) size ('kringle iv' repeat) polymorphism and a pentanucleotide repeat polymorphism influences Lp(a) lipoprotein level. *J. Intern. Med.* **247**(1):139–152.

Sachse, C., Brockmoller, J., Bauer, S., and Roots, I., 1997. Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. *Am. J. Hum. Genet.* **60**(2):284–295.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., and Miller, W., 2003. Human-mouse alignments with blastz. *Genome Res.* **13**(1):103–107.

Toribara, N. W., Gum,Jr., J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M., and Kim, Y. S., 1991. MUC-2 human small intestinal mucin gene structure. Repeated arrays and polymorphism. *J. Clin. Invest.* **88**(3):1005–1013.

Uhrberg, M., Valiante, N. M., Shum, B. P., Shilling, H. G., Lienert-Weidenbach, K., Corliss, B., Tyan, D., Lanier, L. L., and Parham, P., 1997. Human diversity in killer cell inhibitory receptor genes. *Immunity* **7**:753–763.