2004

Analysis of Human mRNAs With the Reference Genome Sequence Reveals Potential Errors, Polymorphisms, and RNA Editing

Terrence S. Furey
University of California - Santa Cruz

Mark Diekhans
University of California - Santa Cruz

Yontao Lu
University of California - Santa Cruz

Tina A. Graves
Washington University School of Medicine in St. Louis

Lachlan Oddy
Washington University School of Medicine in St. Louis

See next page for additional authors

Analysis of Human mRNAs With the Reference Genome Sequence Reveals Potential Errors, Polymorphisms, and RNA Editing

Terrence S. Furey, Mark Diekhans, Yontao Lu, et al.

Genome Res. 2004 14: 2034-2040
Access the most recent version at doi:10.1101/gr.2467904

Letter

Analysis of Human mRNAs With the Reference Genome Sequence Reveals Potential Errors, Polymorphisms, and RNA Editing

Terrence S. Furey,[1,4] Mark Diekhans,[1] Yontao Lu,[1] Tina A. Graves,[2] Lachlan Oddy,[2] Jennifer Randall-Maher,[2] LaDeana W. Hillier,[2] Richard K. Wilson,[2] and David Haussler[3]

[1]Center for Biomolecular Science and Engineering, Department of Computer Science, University of California, Santa Cruz, Santa Cruz, California 95064, USA; [2]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA; [3]Howard Hughes Medical Institute, Center for Biomolecular Science and Engineering, Department of Computer Science, University of California, Santa Cruz, Santa Cruz, California 95064, USA

The NCBI Reference Sequence (RefSeq) project and the NIH Mammalian Gene Collection (MGC) together define a set of ~30,000 nonredundant human mRNA sequences with identified coding regions representing 17,000 distinct loci. These high-quality mRNA sequences allow for the identification of transcribed regions in the human genome sequence, and many researchers accept them as the correct representation of each defined gene sequence. Computational comparison of these mRNA sequences and the recently published essentially finished human genome sequence reveals several thousand undocumented nonsynonymous substitution and frame shift discrepancies between the two resources. Additional analysis is undertaken to verify that the euchromatic human genome is sufficiently complete—containing nearly the whole mRNA collection, thus allowing for a comprehensive analysis to be undertaken. Many of the discrepancies will prove to be genuine polymorphisms in the human population, somatic cell genomic variants, or examples of RNA editing. It is observed that the genome sequence variant has significant additional support from other mRNAs and ESTs, almost four times more often than does the mRNA variant, suggesting that the genome sequence is more accurate. In ~15% of these cases, there is substantial support for both variants, suggestive of an undocumented polymorphism. An initial screening against a 24-individual genomic DNA diversity panel verified 60% of a small set of potential single nucleotide polymorphisms from which successful results could be obtained. We also find statistical evidence that a few of these discrepancies are due to RNA editing. Overall, these results suggest that the mRNA collections may contain a substantial number of errors. For current and future mRNA collections, it may be prudent to fully reconcile each genome sequence discrepancy, classifying each as a polymorphism, site of RNA editing or somatic cell variation, or genome sequence error.

[Supplemental material is available online at www.genome.org.]

The production of a high-quality human genome sequence has allowed researchers to begin exploring the genome in new and exciting ways. Gene sequences can now be viewed not simply as isolated and processed mRNA sequences but as complete genomic units with distinct exon/intron structures, regulatory regions, and genomic contexts. The genome sequence is a template on which we can now map minor genetic variations in the human population such as single nucleotide polymorphisms (SNPs) and polymorphic insertions and deletions (indels) of genome sequence. In genes, these can cause subtle changes in the translated amino acid sequence that can profoundly affect how the protein behaves. In biomedical research, a current focus is determining the relationship between gene alleles and phenotypic differences such as susceptibility to disease and response to drug treatment.

The accurate identification of gene sequences within the human genome sequence is only possible due to the existence of high-quality collections of full-length mRNA sequences. Purely computational efforts to accurately and comprehensively identify gene sequences have been unsuccessful in this regard. The GenBank (Benson et al. 2002), EMBL (Kulikova et al. 2004), and DDBJ (Miyazaki et al. 2004) nucleotide databases have been central repositories for mRNA sequences with continual synchronization between them. To help make sense of the vast number of these sequences of varying accuracy, the Reference Sequence (RefSeq) project (Pruitt et al. 2003) was started with the aim of creating a high-quality, nonredundant set of full-length mRNA sequences from GenBank to act as the gold standard for gene sequences. Independently and more recently, the Mammalian Gene Collection (MGC; Strausberg et al. 1999; MGC Program Team 2002; MGC Project Team 2004) began producing a set of high-quality, full-length mRNA sequences based on their collection of cDNA clones. The combined alignments of mRNA sequences from these two collections identify >17,000 distinct gene loci in the human genome sequence.

mRNA sequences have consistently provided the best representation of gene sequences, yet a detailed evaluation of the quality of these mRNAs has not been performed until now. With the human genome sequence now essentially finished, we have a second independent and high-quality resource that can be used for this type of analysis. By aligning RefSeq and MGC mRNA sequences to the genome sequence, discrepancies can be identified and further explored for possible errors as well as for poly-

[4]Corresponding author.
E-MAIL booch@cse.ucsc.edu; FAX (831) 459-4829.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2467904.

14:2034–2040 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

morphisms and sites of RNA editing or somatic cell variation. With nearly the entire genome sequence, we can be confident that our alignments correctly reflect the true origin of the mRNA sequence, which is critical to this type of detailed analysis. The International Human Genome Sequencing Consortium (International Human Genome Sequencing Consortium 2001) boasts that the genome base pair error rate is less than one in 10,000 genome-wide (Schmutz et al. 2004). On average, this would amount to less than one incorrect base pair in a typical gene sequence. Even with this high sequencing standard, errors in the genome sequence have arisen due to deletions and other mutations introduced during creation and amplification of the bacteria artificial chromosome (BAC) clones used to facilitate the sequencing of the genome. We found, however, that these deletions along with known unsequenced gaps affect <1% of gene sequences, thus the genome sequence remains an excellent resource of mRNA evaluation.

cDNA cloning and mRNA sequencing typically involve multiple rounds of cloning and amplifying DNA sequence that has been reverse transcribed from an mRNA transcript (Yu et al. 1997; Gunaratne et al. 2003). Reverse transcriptase (RT) is a known source of error with error rate estimates ranging between one in 1000 to one in 15,000 (Stapelton et al. 2002). In addition, the stringent sequencing standard enforced while sequencing the genome has not been applied to mRNA sequences submitted to GenBank. In fact, a significant number of RefSeq sequences are based on mRNAs sequenced in the early to mid 1990s before more recent advances in sequencing technology. Sequencing of MGC cDNA clones is more recent, and the quality of these sequences is better on average than that of those in RefSeq and is comparable to that of the genome sequence. Although many errors can be detected when the mRNA sequence contains a poor open reading frame (ORF), it is hard to distinguish cloning or sequencing errors from true polymorphism. For the remainder of this article, we refer to the existence of potential cDNA cloning errors or mRNA sequencing errors as simply potential mRNA sequence errors, as it cannot be determined at what stage the errors may have occurred.

We present here an assessment of the accuracy of full-length mRNA sequences in the RefSeq and MGC collections based on an in-depth analysis of alignments to the July 2003 release of the human genome sequence using BLAT (Kent 2002). In parallel, the MGC Project Team has performed an analysis of their cDNA clones (MGC Project Team 2004) with comparable results. We have concentrated our analysis on the annotated protein coding region (CDS) of each mRNA. We identify all cases in which there is a disagreement in the identity of single bases at a given position with further analysis of those that cause amino acid changes in the resulting protein. For each of these single base discrepancies, we initially compare them against documented SNPs in the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/, release 119). For those not present, we survey all other human and nonhuman mRNA and EST sequences available for the locus to determine whether there is support for the mRNA sequence variant and/or the genome sequence variant. Although we can never completely rule out the possibility of rare polymorphisms, somatic cell genomic variants, or instances of RNA editing, significant support for one sequence and little or no support for the other does suggest a sequence error may have occurred. Additional analysis of ratios of synonymous to nonsynonymous substitutions support this assertion. This ratio is similar to the one reported by the MGC Project Team (MGC Project Team 2004). We also investigate indels in the mRNA sequence compared with the genome sequence in a similar way. We present combined results for mRNAs from both the RefSeq and MGC collections, although it is important to note that although over half of the sequences originated from the MGC collection, only one third of the total discrepancies come from MGC produced mRNAs.

Based on these analyses, we find that when a discrepancy is detected, the genome sequence variant is supported by at least three independent human transcript sequences over three times as often as the mRNA sequence. For half of the mismatches and 40% of the indels, the mRNA has no independent transcript support for the region in question, although this is true of the genome sequence only 10% of the time. We believe that this suggests that the mRNA sequence collections do contain errors, and that this type of analysis can point to specific instances that require further investigation. In fact, we believe that all current and future mRNA sequences should be fully reconciled with the genome sequence in order to ensure the accuracy of these sequences and to identify potential errors in the genome sequence. We must admit that our analysis depends on the accurate alignment of transcript sequences, and it is highly probable that there are errors in a small number of these BLAT alignments. A comparison of alignments with those produced by SPIDEY (Wheelan et al. 2001) and BLAST (Altschul et al. 1990) indicates that ~1.5% of these BLAT alignments might be slightly incorrect but that BLAT produces the best alignments overall (data not shown).

These analyses also point to potential, not previously documented instances of coding polymorphisms. We identify >2000 unique bases in coding regions that we believe are potential SNPs with three or more independent transcripts supporting each variant. We tested 25 of these against a panel of 24 diverse genomic DNA samples, and of the 20 that could be successfully assayed, 12 were found to be polymorphic. The inability to confirm the other eight cases does not indicate that they are not polymorphic as they may be rarer polymorphisms simply not found in the panel of 24 samples.

Potential polymorphic indels were also identified and tested in a similar way. Of six potential nonframeshift indels that could be successfully assayed, three instances could be shown to be polymorphic. Of 14 potential frameshift indels that could be successfully assayed, only one could be verified as polymorphic. In 12 of the remaining 13 cases, there is only evidence for the genomic variant. This might indicate sites particularly prone to error during the cDNA cloning and/or mRNA sequencing process, or it may point to sites of posttranscriptional RNA editing. The most common form of mRNA editing, however, involves the modification of a single base and not the insertion or deletion of bases. We do also find statistical evidence of this more common form of RNA editing as part of this analysis.

RESULTS

Each of the 30,820 mRNA sequences from the RefSeq and MGC collections was aligned to the genome sequence by using BLAT. mRNA sequences with overlapping annotated coding regions (CDS) were clustered into gene loci, and the corresponding 17,019 loci were classified as described in the Methods section and summarized in Table 1. This does reveal that some loci are not completely represented in the genome sequence, primarily due to sequence gaps, but only 99 loci (<1%) are affected (see Supplemental Tables 6, 7). The completeness of the genome sequence is further supported by our ability to find 99.7% of ~8000 sequence-tagged site markers from the Genethon (Dib et al. 1996), Marshfield (Broman et al. 1998), and deCODE (Kong et al. 2002) genome-wide genetic maps.

A detailed analysis was performed on alignments of the annotated CDS of the 30,631 unique mRNA sequences determined to be found with 95% of the mRNA sequence aligned at >98% base pair identity. These alignments involve ~43 Mb of sequence. In 17,758 (58.0%) instances, the CDS sequence aligned perfectly

Table 1. Gene Loci Coverage

Category	Loci	Sequences
Found	16,920 (99.4%)	30,663 (99.5%)
Partially found	81 (0.5%)	127 (0.4%)
Missing	18 (0.1%)	30 (0.1%)
Total	17,019	30,820

Full-length mRNA sequences are aligned to the genome sequence and classified as described in the Methods section. Detailed in this table are the number of mRNA sequences and the corresponding number of loci the sequences represent that are found, partially found, or missing in the genome sequence.

to the genome sequence with no mismatches, insertions, or deletions. If mismatches at documented locations of SNPs as identified in dbSNP are ignored, an additional 3665 (12.0%) instances were in complete agreement. This leaves 9208 (30%) discrepant alignments, 8585 with one or more mismatched base pairs, and 1479 with indels in the mRNA sequence compared with the genome sequence (some sequences contain both mismatches and indels).

Single Base Pair Mismatches

We find 19,470 single base disagreements not documented in dbSNP. For each of these cases, all human and nonhuman mRNA and EST sequences also aligned and covering the base in question were compared with both the mRNA and genome sequences. (An effort was made to not count mRNAs or ESTs derived from the same cDNA clone used to produce the mRNA sequence in question, or to double count mRNAs and ESTs derived from the same clone. This was done based on clone information in the GenBank records.) For each of the 19,470 discrepancies, we tallied the support from these other transcript sequences. We conservatively require three or more transcript sequences to label a variant well-supported. The results are summarized in Table 2. The genome sequence is much better supported in general than is the mRNA sequence based on this criteria. In 15% of the cases, there is significant human transcript support for both sequences, suggesting the existence of a coding SNP not in dbSNP. These are discussed in more detail below.

Nonsynonymous Codon Substitutions

Of special interest are those mismatches that result in a nonsynonymous amino acid substitution. We find 11,041 instances of these substitutions in 6094 alignments that are not currently documented in dbSNP. Table 3 summarizes the human and nonhuman transcript support for these instances. As is expected based on the above mismatch analysis, there is significantly more support for the genome sequence variant.

If you consider all possible nucleotide changes in every possible codon, about three times as many would result in a nonsynonymous amino acid substitution as would a synonymous substitution. Because of selective pressure, however, you would expect to find a higher ratio of synonymous to nonsynonymous substitutions (S/NS) than would be expected by a random mutation model. We looked at >10,000 codon substitutions based on the RefSeq and MGC mRNA alignments that were also documented in dbSNP and found the the S/NS ratio was 1.35, as shown in Table 4. Similarly, we looked at ~2700 codon substitutions not documented in dbSNP but with significant support for both the genome and mRNA variants. The S/NS ratio for this set was 1.27, providing additional confidence that this set accurately identifies undocumented polymorphism. In contrast, the 11,000 substitutions found based on the mRNA alignments that are not in dbSNP and with either the genome variant or the mRNA variant that has no human transcript support have a S/NS ratio of 0.58. This suggests that the majority of these point to a genome or mRNA sequence error.

We also looked at S/NS ratios in the RIKEN mouse mRNA collection (Okazaki et al. 2002) identified through their alignment to the October 2003 release of the mouse genome sequence (Mouse Genome Sequencing Consortium 2002). Considering only those mRNAs aligning to finished sequence, we find a S/NS ratio of 1.77 for instances in which both the mRNA and genome sequence were supported by three additional mouse mRNA or EST sequences. Those with only support for either the mRNA or genome sequence variant had a S/NS ratio of 0.36.

RNA Editing

We surveyed 9660 base pair mismatch cases that were not documented in dbSNP and in which the mRNA or genome sequence variant had at least three supporting human transcripts, whereas the opposing variant had no support. We believe that these instances consist primarily of cases of sequence error. The most

Table 2. CDS Mismatch Support

	≥3 Human mRNA/EST	≥3 Non-human mRNA/EST	No support
Genome only	12,000 (61.6%)	9686 (49.7%)	1249 (6.4%)
mRNA only	1339 (6.9%)	1863 (9.6%)	10,270 (52.7%)
Genome and mRNA	2925 (15.0%)	3822 (19.6%)	824 (4.2%)
Neither	3206 (16.5%)	4099 (21.1%)	N/A

Analysis of 19,470 mismatched base pairs in alignments of CDS regions of full-length mRNAs to the human genome sequence. Sequences are compared against other human and nonhuman mRNA and EST sequences that align at this position. Reported are the number of instances in which three or more other human (column 2) or nonhuman (column 3) mRNA or EST sequences agree with each sequence. The last column indicates how many have no other human or nonhuman mRNA or EST support at all. The first row shows the number of times only the genome has the corresponding level of support; the second row, the same for the mRNA. The third row shows how many times both the genome and the mRNA are supported at the level indicated; the last row, when neither have that level of support.

Table 3. Nonsynonymous Substitution Support

	≥3 Human mRNA/EST	≥3 Non-human mRNA/EST	No support
Genome only	7144 (64.7%)	5823 (52.7%)	659 (6.0%)
mRNA only	754 (6.8%)	677 (6.1%)	5737 (52.0%)
Genome and mRNA	1183 (10.7%)	396 (3.6%)	372 (3.4%)
Neither	1960 (17.8%)	4145 (37.5%)	N/A

Analysis of 11,041 nonsynonymous substitutions in alignments of CDS regions of full-length mRNAs to the human genome sequence. Sequences are compared against other human and nonhuman mRNA and EST sequences that align at this position. Reported are the number of instances in which three or more other human (column 2) or nonhuman (column 3) mRNA or EST sequences agree with each. The last column indicates how many have no other human or nonhuman mRNA or EST support at all. The first row shows the number of times only the genome has the corresponding level of support; the second row, the same for the mRNA. The third row shows how many times both the genome and the mRNA are supported at the level indicated; the last row, when neither have that level of support.

Table 4. Ratio of Synonymous to Nonsynonymous Substitutions

	Syn subs	Nonsyn subs	Total subs	Ratio S/NS
All substitutions	13,465	15,494	28,961	0.87
SNP substitution	5998	4453	10,451	1.35
Non-SNP substitutions	7467	11,041	18,508	0.68
Both well-supported	1505	1183	2688	1.27
Support mRNA only	548	752	1300	0.73
Support genome only	3499	6215	9714	0.56

Ratios of synonymous to nonsynonymous substitutions detected by mRNA alignments to the human genome sequence are shown. Substitutions are classified into SNPs or non-SNPs based on inclusion in the dbSNP database. Well-supported substitutions imply that there are at least three additional transcript sequences supporting both the genome and mRNA sequence variants. For the last two cases in which only the mRNA or genome sequence variant is supported, we only require a single additional supporting transcript sequence.

common sequence errors are those involving transitions between C and T, or between A and G, due to the similarity in chemical structure and the higher rate of mutation between these base pairs. As expected, our set was dominated by these types of mismatches, as shown in Table 5. However, some of these with only support for the mRNA sequence may be instances of RNA editing.

RNA editing consists of modifying pre-mRNA sequences before splicing and translation (Bass 2002). In some instances, this editing is performed on the majority if not all of the pre-mRNAs transcribed at a particular locus. For these cases, it is difficult to distinguish RNA editing from probable sequence errors as there would be a discrepancy between the genome and mRNA sequences and only additional transcript support for the mRNA sequence. Nevertheless, one might see some hint of RNA editing in the statistics of the observed changes. Therefore, we looked specifically at the 796 of the 9660 mismatch cases that show only transcript support for the mRNA sequence. Table 5 also shows the results for this subset.

It is immediately obvious that this subset is enriched for mismatches in which the genome sequence has an A at a location where the mRNA and all other transcripts report a G. The best known case of RNA editing involves the deamination of an A to inosine (I), which is read as a G and paired with C during translation (Gerber and Keller 2001; Bass 2002). The increase in the proportion of A (genome) to G (mRNA) mismatches suggests that some of these may not be sequence errors, but cases of RNA editing. A well-known case of RNA editing involves the glutamate receptor *GluR2* gene (RefSeq NM_000826) for which this process is critically regulated in the brain (Sommer et al. 1991; Paschen et al. 1994; Kawahara et al. 2003). The edited base for this mRNA sequence is included in our set of 796 transitions described above, evidence that this set is enriched for instances of RNA editing.

Surprisingly, the asymmetric case of T in the genome versus a C in the mRNA and all other transcripts is also found much more often, proportionally, in the subset. Although this type of RNA editing has been reported and is hypothesized to possibly involve a "trans"-amination reaction (Gerber and Keller 2001), it is not as well-known or characterized as the above. Equally surprising is the lack of enrichment for the case of C in the genome versus a T in the mRNA, as the deamination of C resulting in a T is a well-known type of RNA editing (Gerber and Keller 2001). It is possible, however, that loci undergoing this type of edit also produce a significant number of unedited transcripts, thus support for the genome sequence would have been seen and these would have been excluded from our set.

Table 5. Unsupported Mismatches

Genome base	mRNA base	All support	No Genome support
A	G	1424 (14.7%)	186 (23.3%)
G	A	1675 (17.3%)	90 (11.3%)
C	T	1179 (12.2%)	74 (9.3%)
T	C	1358 (14.1%)	168 (21.1%)

This table shows that most common mismatched bases between the mRNA and genome sequences in which one is supported by at least three human transcripts, the other by none. The third column reports results from all 9660 instances, and the fourth column shows results from the 796 of these with only support for the mRNA sequence and none for the genome sequence.

Indels

We define insertions in an mRNA sequence as any stretch of unaligned bases in the coding region of the mRNA sequence where the surrounding mRNA sequence aligns without a gap in the genome sequence. Deletions are similarly defined as small stretches of <30 unaligned bases in the genome sequence that lack a standard splice signal and for which the surrounding genome sequence aligns without a gap to the mRNA sequence. Longer deletions are more likely to be cases of nonstandard introns, or introns in which sequence errors have affected the splice signal. There are 2612 indels in 1479 alignments based on these definitions. Of these 2612 indels, 381 involve three bases or a multiple of three bases that would not result in a frameshift and conceivably are polymorphisms.

Of the remaining 2231 cases, 2008 are single base indels, the majority of which are most probably due to sequence errors. As with mismatches, these frameshift indels are evaluated by using other transcript sequences, and the results are summarized in Table 6. Once again, we see significantly more support for the genome sequence. Similar results have been obtained by using the draft chimpanzee genome sequence as support for either the mRNA sequence or the human genome sequence (T. Mikkelsen, pers. comm.).

Table 6. Insertions and Deletions in Coding Regions

	≥3 Human mRNA/EST	≥3 Non-human mRNA/EST	No support
Genome only	1345 (60.3%)	1155 (51.8%)	115 (5.2%)
mRNA only	204 (9.1%)	145 (6.5%)	767 (34.4%)
Genome and mRNA	275 (12.3%)	266 (11.9%)	94 (4.2%)
Neither	407 (18.2%)	665 (29.8%)	NA

Analysis of 2231 frameshift insertions and deletions in alignments of CDS regions of full-length mRNAs to the human genome sequence. Sequences are compared against other human and non-human mRNA and EST sequences that align at this position. Reported are the number of instances in which three or more other human (column 2) or nonhuman (column 3) mRNA or EST sequences agree with each. The last column indicates how many have no other human or non-human mRNA or EST support at all. The first row shows the number of times only the genome has the corresponding level of support; the second row, the same for the mRNA. The third row shows how many times both the genome and the mRNA are supported at the level indicated; the last row, when neither have that level of support.

Potential Polymorphism

Some discrepancies that we have identified will be explained by polymorphisms in the population. In Tables 2, 3, and 6, 10% to 15% of the identified discrepancies have significant human transcript support for both the mRNA and genome sequence variants. This suggests that these are potential instances of polymorphism. Table 2 does not include 10,523 instances of mismatches due to documented SNPs found in dbSNP. Of these, 4923 (47%) have three or more transcripts supporting both the mRNA and genome sequence variants. Also, 7848 of all mismatches have significant support for both the mRNA and genome sequence variants, so 63% (4923/7848) of these have already been shown to be polymorphic. Together, these support our hypothesis that significant levels of transcript support for both the mRNA and genome sequence variants are a good indicator of polymorphism.

As shown in Table 2, there are 2925 instances of mRNA sequence mismatches with this level of transcript support that are not in dbSNP. Supplemental Table 1 lists locations of these potential SNPs with accessions of transcript sequences that support the alternative variants. Supplemental Table 2 documents the subset of these that are potential nonsynonymous SNPs, providing their location, supporting evidence, and the resulting amino acid substitution.

These 2295 instances identify 2123 unique sites of potential polymorphism (multiple mRNA sequences can identify the same site). By using a panel of 24 diverse genomic DNA samples, we tested 25 of these locations for polymorphism as described in the Methods section. A complete set of results for all experiments using the diversity panel can be found in Supplemental Table 5. For five cases, the experiment failed. Of the remaining 20 cases that succeeded, 12 were confirmed to be polymorphic in this small set of individuals. Of the remaining eight instances, only the genomic variant was found in six cases, whereas in the other two only the mRNA variant was found. This does not eliminate the possibility that these eight instances are polymorphic. A further investigation into one of these cases (mRNA BC012449.1, base 311) shows that three BAC clones from the RP11 library had been sequenced that contained the site in question, and two contained the genome sequence variant, whereas the other contained the mRNA sequence variant. Thus, this suggests that the site is polymorphic but may be a rare polymorphism not found in the panel of 24 genomic DNA samples.

As mentioned previously, we found 381 instances of indels that did not result in a frameshift in the coding sequence and are not included in Table 6. Supplemental Table 3 lists locations of these in-frame coding indels. Of these, 100 have significant support for both the mRNA and genome sequence variants, identifying 80 unique locations of potential polymorphism. In the same manner as above, we tested seven of these instances with the diversity panel. In three of the six cases that were successful, the site was determined to be polymorphic. For the remaining three, all samples in the diversity panel agreed with the genome sequence variant.

In Table 6, we report 275 instances of frameshift indels with significant support for both the mRNA and genome sequence variants. Supplemental Table 4 lists the locations of these. After eliminating redundancy, we have 239 unique sites. We tested 16 of these with the diversity panel, with 14 giving successful results. Only one site, a one base deletion near the end of the coding region in BC020779 (NM_23666), with a product that is thought to be a phosphatidylethanolamine binding protein, was found to be polymorphic. The translation of BC020779 results in a protein sequence four amino acids smaller than that of the translated genome-defined variant (also represented by mRNA AY037148), and changes the identity of the last four amino acids in this shorter allele compared with the last eight in the longer one. In 12 of the remaining 13 cases, only the genomic variant was found, whereas for the last case, only the mRNA variant was present.

It is interesting that of the 16 indels (three nonframeshift, 13 frameshift) that were not confirmed polymorphisms, all samples in the diversity panel agreed with the genome sequence except in one case. This high number of mRNA discrepancies may just be a random result of RT errors, as suggested in the substitutions analysis above, or there may be something particular about these sites that causes a higher rate of discrepancy. Further investigation is required.

DISCUSSION

The essentially finished genome sequence provides a complete and accurate representation for nearly all gene loci. The RefSeq and MGC full-length mRNA collections are of high quality and are necessary for determining precisely the location of each locus in the genome, and also help define gene sequence variants derived from polymorphisms and alternative splicing. Purely computational methods have failed to identify these gene structures with both high sensitivity and specificity, and experimental evidence provided by mRNA and EST sequences is crucial. But, the quality of mRNA sequences can be highly variable, depending on the quality of the cDNA clone, the method of sequencing, and the effort made to ensure its accuracy.

The genome sequence provides a representative sequence for the vast majority gene loci. Our analysis suggests that this genome sequence is more accurate than are mRNA sequences based on support of independent transcript sequences at sites where the sequences disagree. It is important to note that these discrepancies are still rare when considering the total amount of sequence analyzed. If we exclude documented polymorphisms and probable polymorphisms we detect, then there are only approximately four discrepancies for every 10,000 bases. This is definitely an upper bound on any sequence error rate, as many of these will be cases of polymorphism, RNA editing, or somatic cell-specific genomic variation. Because of the importance of accurate gene sequences, we believe that an investigation and characterization of all discrepancies is necessary to ensure the quality of mRNA collections.

In addition, the investigation of these discrepancies can lead to meaningful biological discovery such as new cases of coding polymorphisms and RNA editing. Our initial investigation of potential polymorphisms suggested by this analysis suggests that a majority of our predicted coding SNPs are real. In addition to further validation of polymorphisms, it would be interesting to search more closely for new cases of RNA editing. Currently characterized sites reveal that the editing mechanism depends on the formation of a hairpin loop by the RNA surrounding the edited base (Gerber and Keller 2001; Bass 2002). Using our current analysis to look for potential sites combined with a further computational search for this RNA structure should lead to good candidates that can then be evaluated experimentally.

METHODS

RefSeq and MGC Full-Length mRNA Sequences

Full-length mRNA sequences from the RefSeq project (Pruitt and Maglott, 2001) and/or the MGC (Strausberg et al. 1999; MGC Program Team 2002) were obtained from the http://www.ncbi.nlm.nih.gov/RefSeq/ and GenBank, respectively. We use mRNAs that are present in the collections as of May 16, 2003. The actual sequences and coding region annotations for these mRNAs were downloaded August 1, 2003. We have discarded some sequences as detailed in Supplemental Text 1, leaving

30,820 nonredundant mRNA sequences representing 17,019 distinct loci. Many of these discarded sequences have been among those recently removed and/or updated in the RefSeq and MGC collections, with a significant number of these changes due to the research presented here. For this analysis, we define a loci as consisting of a set of mRNA sequences with one or more overlapping bases from each of their annotated coding regions.

Human Genome Sequence

The July 2003 human genome sequence (NCBI build34) consists of ordered and oriented sequence for each of the 22 autosomes and two sex chromosomes using only finished sequence. In addition, sequence that is finished but represents a different haplotype, and sequence that is in a draft state and not yet finished are included in build34. All sequence in build34 is used this analysis.

Placement of mRNA Clone Sequences

Locations of full-length mRNA sequences are determined using BLAT version 24 with parameters −q = rna −trimHardA −fine −ooc = 11.ooc. Resulting alignments are filtered to report only the best alignments for each mRNA, requiring at least 98% base pair identity. As mentioned above, a small subset of alignments with less than but close to 98% base pair identity are considered found with the poor base pair identity attributed to a difference in the haplotype from which the mRNA sequence and genome sequence were derived.

Determination and Classification of Gene Loci

The mRNA sequences were clustered into gene loci where each locus contained all mRNA sequences with aligned annotated coding region that overlapped on the same genomic strand. For mRNAs that lacked a genome alignment, loci were defined by using information in the LocusLink resource (Pruitt and Maglott 2001) and the Stanford SOURCE database (Diehn et al. 2003). The following summarizes how each of the loci were classified into found, partially found, and missing categories. More complete details on the following classification can be found in Supplemental Text 2.

A gene loci is classified as found if the mRNA sequences defining that loci could be aligned by BLAT at 98% base pair identity >95% of the mRNA sequence. For 112 loci, the mRNA sequences could be aligned equally well at multiple locations in the genome, and these are found to correspond to recent segmental duplications (Bailey et al. 2002). Another 33 mRNA sequences were deemed found even though they did not meet the above criteria. This group is comprised primarily of immunoglobulin and major histocompatibility mRNA sequences which are known to be highly polymorphic in the population.

Of the remaining 99 loci, mRNA sequences from 81 could be partially aligned by BLAT at 98% base pair identity. Supplemental Table 7 provides a listing and locations for each of these 81 loci. In 46 of these 81 cases, the partial alignment is due to an incompleteness in the genome sequence. For the other 35 partial alignment cases, there is evidence that a deletion or misassembly prevents a full alignment. For the last 18 loci not considered found, no alignment for the mRNAs could be made. A full list of these 18 loci are provided in Supplemental Table 6. Of these, 12 can be mapped to existing unsequenced gaps, four to unsequenced heterochromatic regions, one is missing due to an assembly error, and the last has not been mapped.

Assaying Potential Polymorphisms

To investigate whether discrepancies between mRNA and the genome sequences are due to polymorphisms, PCR products were generated and resequenced, and the potentially polymorphic bases examined in the DNA from a panel of 24 ethnically diverse individuals (The International Human SNP Working Group 2001).

ACKNOWLEDGMENTS

We would like to thank Francis Collins, Bob Waterston, Jim Kent, Greg Schuler, Daniela Gerhard, David Jaffe, the chromosome coordinators, and all the members of the International Human Genome Sequencing Consortium for their hard work in creating the human genome sequence and providing feedback to our analysis; David Lipman and Lukas Wagner for their suggestion to review the synonymous/nonsynonymous substitution ratios; Tarjei Mikkelsen for help with the indel analysis; Jeremy Schmutz for discussions on mRNA and genome sequencing; Kim Pruitt and Barbara Ruef at RefSeq for their many and timely responses to our questions; Gill Bejerano for help with the RNA editing analysis; Chimpanzee Genome Sequencing Consortium for the availability of the draft chimpanzee sequence.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

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:** 1003–1007.

Bass, B. 2002. RNA editing by adenosine deaminases that act on RNA. *Ann. Rev. Biochem.* **71:** 817–846.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B., and Wheeler, D. 2002. GenBank. *Nucleic Acids Res.* **30:** 17–20.

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L., and Weber, J.L. 1998. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63:** 861–869.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E. et al. 1996. A comprehensive genetic map of the human genome based on 5265 microsatellites. *Nature* **380:** 152–154.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J., Hernandez-Boussard, T., Rees, C., Cherry, J., Botstein, D., Brown, P. et al. 2003. SOURCE: A unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31:** 219–223.

Gerber, A.P. and Keller, W. 2001. RNA editing by base deamination: More enzymes, more targets, new mysteries. *Trends Biochem. Sci.* **26:** 376–384.

Gunaratne, P., Wu, J., Garcia, A., Hulyk, S., Worley, K., Margolin, J., and Gibbs, R. 2003. Concatenation cDNA sequencing for transcriptome analysis. *C. R. Biol.* **326:** 971–977.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Kawahara, Y., Ito, K., Sun, H., Kanazawa, I., and Kwak, S. 2003. Low editing efficiency of GluR2 mRNA is associated with a low relative abundance of ADAR2 mRNA in white matter of normal human brain. *Eur. J. Neurosci.* **18:** 23–33.

Kent, W.J. 2002. BLAT: The BLAST-like alignment tool. *Genome Res.* **11:** 1541–1548.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhart, R. et al. 2004. The EMBL mucleotide sequence database. *Nucleic Acids Res.* **32:** D27–D30.

MGC Program Team. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899–16903.

MGC Project Team. 2004. The status, quality, and expansion of the NIH full-length cDNA project: The mammalian gene collection (MGC). *Genome Res.* (this issue).

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. 2004. DDBJ in the stream of various biological data. *Nucleic Acids Res.* **32:** D32–D34.

Mouse Genome Sequencing Consortium 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. et al. 2002. Analysis of

the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Paschen, W., Hedreen, J., and Ross, C. 1994. RNA editing of the glutamate receptor subunits GluR2 and GluR6 in human brain tissue. *J. Neurochem.* **63:** 1596–1602.
Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.
Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2003. NCBI Reference sequence project: Update and current status. *Nucleic Acids Res.* **31:** 34–37.
Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan Y.M., Denys, M., et al. 2004. Quality assessment of the human genome sequence. *Nature* **429:** 365–368
Sommer, B., Kohler, M., Sprengel, R., and Seeburg, P., 1991. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67:** 11–19.
Stapelton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S. et al. 2002. A *Drosophila* full-length cDNA resource. *Genome Biol.* **3:** RESEARCH0080.1–0080.8.
Strausberg, R.L., Feingold, E.A., and Collins, F.S. 1999. The mammalian gene collection. *Science* **286:** 455–457.
Wheelan, S.J., Church, D.M., and Ostell, J.M. 2001. Spidey: A tool for mRNA-to-genomic alignments. *Genome Res.* **11:** 1952–1957.
Yu, W., Andersson, B., Worley, K., Muzny, D., Ding, Y., Liu, W., Ricafrente, J., Wentland, M., Lennon, G., and Gibbs, R. 1997. Large-scale concatenation cDNA sequencing. *Genome Res.* **7:** 353–358.

WEB SITE REFERENCES

http://www.ncbi.nlm.nih.gov/SNP/; dbSNP database.
http://www.ncbi.nlm.nih.gov/RefSeq/; RefSeq project.

Received February 16, 2004; accepted in revised form May 27, 2004.