# Appendix A: Annotation on human chromosome 13q32-33 by DNannotator

## *Target gDNA sequence:*

(available at http://sky.bsd.uchicago.edu/example_data/gDNA-n-gbheader/)

Sequences covering but not limited to the region between D13S122 and D13S779:

1. NCBI30: from NCBI's assemblies build 30 (NT_009952.10, 25 Mb)
2. TA: Our manually assembled sequence of the region (17 Mb)
3. GB-chr13: Genome Browser November 2002 freeze chromosome 13.

## *Annotation Source data:*

(available at http://sky.bsd.uchicago.edu/example_data/source_data/denovo/)

1. 49 cDNA sequences
2. 548 primers
3. 1600 SNPs selected from public databases, 157 SNPs or insertions/deletions identified in our lab
4. STSs from NCBI's UniSTS

## 1. *De novo* Annotation

Table 1 presents the summary of *de novo* annotation of three gDNA assemblies.

All 49 transcripts were recognized and annotated in all different assemblies. All of the annotated exons show consistent orientation and size in the three assemblies. Five extremely short exons (<10 bp) are considered un-reliable calls from Sim4, and a few transcripts with more than 10 bp are not covered by an exon report. These exons and transcripts were labeled with warning messages by DNannotator. Without the problem of calling un-reliable short exons, BLAT-based exon annotation called fewer exons than Sim4. Most of the exons defined by Sim4 and BLAT are consistent with a few exceptions: about 30 exons (accounts for ~5% of the total exons) are called differently by these two methods.

In the case of SNP mapping, using the default parameters (minimum size of valid match of 50 bp, minimum percentage of identities of 95%, maximum length difference between query SNP and BLAST match of 10 bp), only one false positive mapping, rs2009772, was detected by report of duplicated mapping if the second filter (which keeps only best sequence alignment) was turned off.

This SNP has a homologous sequence (>400 bp 95%, which is above default setting of the first filter) in this region. But once this second filter is turned on, no false positive mapping is produced. Five short insertion/deletion markers show false negative results in all 3 assemblies due to excessive gaps introduced in the BLAST matches. Two more SNPs (rs2390760, rs873447) failed annotation in TA due to lower quality gDNA sequence in that specific region in the assembly.

More than 400 STSs were annotated by DNannotator. Because of size differences among the gDNA sequences, the amount of mapped STS varies among assemblies. Some duplicated annotations were observed. Six of them (D13S158, D13S174, D13S278, D13S281, D13S286, D13S128) were caused by redundant records in the UniSTS database. Two distinct markers mapped to different places (1.6 Kb away) on 13q are both named D13S128. Since UniSTS data has been pre-computed in Genome Browser map, we did not provide STS mapping for Genome Browser map.

In the case of mapping primers, duplication of mapping could be a major problem as very short sequence is used for querying. Twelve out of 548 primers were found to be part of repeat sequence, especially in *Alu*, by the DNannotator utility "screen primer for repeats." Four of the 12 primers have more than 100 identical copies in NCBI30. These primers in repeats were excluded for mapping. With the default parameters of mapping only 100% identical primer sequence, less than 40 primers could not be mapped into assemblies. It turns out that most of them (32) were designed for amplification on cDNA, BAC clone vectors, genes on other chromosomes, or were modified by adding extra sequence tails. Five primers (CC-ex15-tc-F, cc-ex20-21F, cc-ex25R, cc-ex27-226SR, exon15-tc-f; detail information refer to source data deposit at beginning of this appdendix) could not be mapped in all NCBI's assemblies NCBI30 but were mapped into TA. One primer (AL162717-SS1302596-R) is in the contrary situation. Further analysis shows that these 6 primers are located in regions with discrepant nucleotides among assemblies, and any primer with a single base difference from the target was rejected from annotation. Two primers (TP-RP6-F, TP-RP6-F-X) were mapped to more than one location in this ~ 20 Mb region because the short primer sequences were duplicated in a non-repeat context. With a much longer gDNA sequence, chromosome 13 sequence presented two more primers mapped to multiple locations.

## 2. Annotation migration

After using DNannotator's utility to merge all annotation with sequence data, all 3268 features annotated in TA created by DNannotator, including 1751 SNPs, 513 primer, 556 exons, 448 STSs, were migrated into NCBI30 by the annotation migration function of DNannotator.

### 3. Comparison between annotation migration and *de novo* annotation

The annotation migration results were compared to results of *de novo* annotation in both NCBI30 and TA. The 5 primers (CC-ex15-tc-F etc.), which failed in *de novo* annotation of NCBI30 but succeeded in annotation of TA were transferred to NCBI30 successfully and correctly, since 80 bp of flanking sequence were added for annotation and a lower stringency filter was used. A small percentage (<4% by default) of polymorphism in a ~100 bp sequence is acceptable for annotation migration because the flanking sequences of a primer can be found in the original annotation. This is different from *de novo* mapping of short primers, in which no sequence context of primers was provided in source data. In other words, annotation migration of primer features has fewer false negatives than *de novo* annotation of primers, because more sequence and lower filter stringency is used. All duplicated primer annotations were faithfully transferred into new annotation.

Three STSs (RH44801, RH92898, WI-10746), two primers (CC-EX39-CA-F, EXON39-CA-F) and one exon (SLC10A2_EXON5) failed to be transferred. Further analyses showed that the flanking sequences of the two primers contain highly polymorphic (TG) repeats, which prevented the BMB annotation. The failed exon contains many degenerate nucleotide codes in TA assembly created by Sequencher, which can be caused by either low quality of original sequence used to make assembly or real polymorphisms, which makes the BLAST-match similarity fall below the annotation filter. Failure of transfer of 3 STSs is due to the polymorphism or sequence error at 25 bp end sequences of the annotated STS, as DNannotator uses these sequences as primers for e-PCR annotation.

*De novo* annotation of NCBI30 has 157 additional STSs mapped to regions beyond the corresponding range of TA (NCBI30 has 8 Mb more sequence than TA), and one more primer (AL162717-SS1302596-R) mapped to the region comparing to annotation migrated from TA. The primer (AL162717-SS1302596-R) failed at the *de novo* annotation of TA because of one base difference at the primer binding site in TA.

In total, 13 elements showed mapping differences between annotations created *de novo* and by annotation migration to NCBI30. All of them were mapped in one way but not in another. Among them, 5 primers as indicated above failed annotation in NCBI30, but they were successfully annotated in TA and transferred to NCBI30. On the other hand, one primer (AL162717-ss1302596-R), two SNPs (rs2390760, rs873447) and two STSs (D13S914**,** SHGC-83034) failed at *de novo* annotation in TA; thus, they are not part of the source for annotation migration. Two primers (CC-ex39-ca-F, exon39-ca-f) and one exon (SLC10A2_exon5) failed at the annotation migration step. Except for the *de novo* annotation of STS, all annotation failures were reported by DNannotator.

**4. Comparison of DNannotator's annotation to public annotation**

DNannotator used selected local source data to create 3409 features on NCBI30. These data were compared with annotations in NT_009952.10 (same sequence as NCBI30) created by NCBI, which contains 20749 SNPs, 686 STSs and 1272 exons of 180 genes. 2619 features were shared in these two data sets, including 592 STSs, 1587 SNPs and 440 exons, in which 3 STSs and 1 SNP (D13S265, D13S266, D13S278 and rs1253823) are mapped differently. Further analysis shows that SNP rs1253823 is located right at the edge of a CA repeat, which created gaps in a BLAST match. Three STSs have 1 or 4 bp difference for their mapping position which actually would not affect the mapping quality of STS at all, as many STSs don't really have clear-cut boundaries. Exon mapping comparison presents most of the differences. Excluding the exons with warning information by DNannotator, 61 exons are mapped differently. Most of the differences (40 out of 61) are due to difference of the public cDNA sequences from the curated sequences used by DNannotator, especially the differences at either 5' or 3' part of cDNA.

To make a side-by-side comparison for exon annotation, 10 cDNA sequences used by NT_009952.10 were used to do *de novo* annotation by DNannotator. Results were compared to NT_009952.10. In the 170 exons created, 14 exons of 6 genes show discrepant boundaries. Further analysis shows that exon analysis is not as straightforward as it sounds. For example, exon 3 of SLC15A1 gene was defined differently by different methods/resources. In NT_009952.10, it's positioned from 12435777 to 12435858 (BLAT supports this result); Sim4 positions this exon between 12435777 to 12435856; NCBI Map Viewer has it between 12435777 to 12435976. Sim4 located the first 25 bp of gene FARP1 between 11852877 and 11852901 in NCBI30. Map Viewer and BLAT all skipped this exon, and assigned the next exon between 11922627 to 11922822, while Sim4 puts the second exon between 11922629 to 11922822. NT_009952.10 has a different exon 1 between 11852660 and 11852901 but his 200 bp exon could not be identified for its cDNA source. In the ~560 exons defined in 13q32-33, about 5% of them have boundaries defined differently by Sim4 and BLAT. Further evaluation of different exon mapping methods is ongoing.

We also observed that besides the SNPs discovered locally, 13 public SNPs (rs1614963, rs1669233, rs1745012, rs1746967, rs1764781, rs1837970, rs1970476, rs517776, rs531090, rs565851, rs574959, rs649589, rs667739) were mapped by DNannotator but unmapped in NT_009952.10. In this list, only rs1837970, rs1970476 and rs531090 were mapped in Genome Browser. There are 6 STSs mapped by DNannotator that were missed in annotation of NT_009952.10. Four of them (G59749, RH45955, SHGC-59595, CDA0QG0) are mapped to more than two different genomic locations in Genome Browser. The other two (D13S1271 and D13S985) are mapped uniquely to 13q. The reason why NCBI does not map those SNPs and STSs remains unclear.

**Table 1.** ***De novo*** **Annotation on three different primary gDNA assemblies on 13q32-33**

| Assembly | SNPs (by BLAST) | | | | STSs (by e-PCR) | | | Primers (by BLAST) | | | | Exons (Gene) (by Sim4) | | | | Exons (Gene) (by BLAT) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Map | Dup | Fail | Warn | Map | Dup | Fail | Map | Dup | Fail | Warn | Map | Dup | Fail | Warn | Map | Dup | Fail | Warn |
| TA | 1750 | 0 | 7 | 9 | 442 | 6 | - | 502 | 2 | 34 | 0 | 561<br>(49) | 0<br>(0) | 0<br>(0) | 10 | 559<br>(49) | 0<br>(0) | 0<br>(0) | 4 |
| NCBI30 | 1751 | 0 | 6 | 7 | 598 | 6 | - | 498 | 2 | 38 | 0 | 561<br>(49) | 0<br>(0) | 0<br>(0) | 10 | 559<br>(49) | 0<br>(0) | 0<br>(0) | 4 |
| GB-chr13 | 1751 | 0 | 6 | 7 | - | - | - | 498 | 4 | 38 | | - | - | - | - | 558<br>(49) | 0<br>(0) | 0<br>(0) | 4 |
| Total | 1757 | | | | Human UniSTS | | | 536 | | | | (49) | | | | (49) | | | |

NCBI30: from NCBI Human Genome build 30; TA: gDNA assembled locally; GB-chr13: chromosome 13 from November 2002 Freeze at Genome Browser.

Total: number of elements tried. Map: number of unique elements mapped. Duplicates of the same elements do not count. Dup: number of elements duplicated.

Fail: number of elements missed. Warn: number of warning message reported in annotation results