Chapter 5

A Computational Scan for U2-Dependent Introns on Human Chromosome 22

Summary

A computational scan for U2-dependent introns on human chromosome 22 identified 3,199 introns strongly supported by expressed sequences. Of these, 0.7 percent were non-canonical GC-AG introns and the remaining 99.3 percent were canonical GT-AG introns. Approximately 12 percent of introns were involved in an intron truncation/extension alternative splicing event, with roughly equal numbers of these occurring at the donor and acceptor ends of the intron. This large set of confirmed introns should prove to be a useful resource for continued detailed gene annotation.

Introduction

Chromosome 22 was the first human chromosome to be essentially completely sequenced (Dunham *et al.,* 1999) and has been extensively annotated using both computational and experimental approaches. Because of this annotation, chromosome 22 has become the de facto test sequence for a large variety of novel informatics analyses. However, annotation is still not complete. In order to assist in the ongoing annotation of chromosome 22, I have developed an expressed sequence based intron identifier, which pairs strong donor and acceptor signals in an attempt to identify as many U2-dependent introns as possible. Results of a preliminary analysis on chromosome 22 and prospects for performing such an analysis on a genomic scale are discussed.

Materials and Methods

Human U2-dependent introns were identified using a two-step procedure, similar to that used in Chapter 4 to identify U12-dependent introns. Potential donor and branch site signals were identified based on statistical pattern recognition techniques, as implemented in StrataSplice (see section 3.4). A posterior probability threshold value of 1e-3 was selected to balance sensitivity with computational demands. Earlier analysis (see sections 3.2, 3.3) suggests that this included roughly 96 percent of GT donor sites, 76 percent of GC donor sites and 96 percent of AG acceptor sites. From these signals, potential introns (donor/acceptor pairs) of less than 20 kb were generated and expressed sequence evidence was used, as described in detad in Chapter 4, to identify a subset of these potential introns as valid. All scans used the 19 May 2000 'Release 2' build of the chromosome 22 sequence (Dunham *et al.,* unpublished, avadable from **http://www.sanger.ac.uk/HGP/Chr22/).**

Confirmed introns were compared to a collection of known repeats on chromosome 22 generated using RepeatMasker (Smit, A.P.R. & Green, P., unpublished) and introns with at least one splice site within a known repeat were discarded and ignored in later analyses.

Results

The computational scan for U2-dependent introns on chromosome 22 generated roughly 72 million potential introns less than 20 kb in length. 4,719 of these potential introns were strongly confirmed by either mRNA or EST sequences. 1,520 of these were found to have at least one splice site (and usually both) within an annotated repeat element. Removing these 1,520 introns left a total of 3,199 expressed sequence confirmed introns on chromosome 22. Nearly 80 percent of these introns agree precisely with chromosome 22 annotations, and this accounts for approximately 70 percent of previously annotated introns (see Table 5.1). 671 of the identified introns were missing from the chromosome 22 annotations and some of these may represent either alternatively spliced forms of known genes or previously unidentified genes.

	Number of Introns
Chromosome 22 Annotation	3,584
U2-Dependent Intron Finder	3,199
Both Sets	2,528
Annotation only	1,056
Intron Finder only	671

Table 5.1 – Comparison between introns identified in this study and annotation of chromosome 22. The total number of introns identified by the chromosome 22 annotation team and the intron finder described in this chapter are provided. The number of introns found in both sets or exclusively in one set is shown as well. Annotation data (Release 2.3, 6 March 2001) were produced by the Chromosome 22 Gene Annotation Group at the Sanger Centre and were obtained from the World Wide Web at **http://www.sanger.ac.uk/HGP/Chr22** (Dunham *et al.*, unpublished).

3,178, or 99.3 percent, of the introns identified in this study were canonical GT-AG introns and the remaining 0.7 percent were non-canonical GC-AG introns. Although the thresholds used in this analysis were biased slightly toward the inclusion of GT-AG introns, these percentages compare well with previously determined estimates of GC-AG intron frequency (International Human Genome Sequencing Consortium, 2001; Burset *et al.*, 2000).

The 3,199 introns identified on chromosome 22 were searched for intron truncation/extension type alternative splicing events in which one splice site remained the same. Of the 3,001 unique donor sites found in the set of 3,199 introns, 175, or 5.8

percent, were associated with two or more acceptor sites. Similarly, 187 or 6.2 percent, of the 2,996 unique acceptor sites were associated with two or more donor sites. Combining these numbers for individual splices sites suggests that roughly 12 percent of U2-dependent introns on chromosome 22 showed intron truncation/extension type alternative splicing. Most alternatively spliced introns only had one alternative form, but a small number of donor and acceptor sites were found that were used in three or four introns (see Table 5.2).

	Donor Sites		Acceptor Sites	
	Number	Percent	Number	Percent
1 Intron	2,826	94.17	2,809	93.76
2 Introns	154	5.13	173	5.77
3 Introns	19	0.63	12	0.40
4 Introns	2	0.07	2	0.07
Total	3,001		2,996	

Discussion

The study described here shows the utility of large-scale intron identification projects for genome annotation and analysis. Preliminary examination of this data by the chromosome 22 annotation team at the Sanger Centre suggests that this approach has identified a variety of potentially novel introns and has provided additional evidence for a large number of previously annotated gene structures.

However the analysis is computationally quite intensive. For instance the analysis of the roughly 35 Mb chromosome 22 sequence involved the generation of approximately 72 million potential introns, which required roughly 9 GB of storage space. Additionally, the search time, even using the high speed SSAHA search algorithm (Ning, Z., Cox, A.J. and Mullikin,J.C., in press) was significant. Searching the 72 million potential introns took 5.5 days of processor time on a 16 GB machme. Furthermore, such an analysis would ideally be performed using more inclusive thresholds, but this would lead to even more significant requirements in terms of disk space and processor time. Because the disk space is needed only transiently, the search time is the key factor determining whether or not such an analysis could be performed on a genomic scale. As analysis of the complete human genome using the same search parameters would take over a year on a high memory machme, the answer at the current time is, realistically, no, at least not as done here.

This approach is much more efficient, however, **if** the expressed sequences are localised within the genome first, allowing the searches to be performed against much smaller databases, and this approach could make such an analysis feasible on a genomic scale. However, it would arguably make more sense to work in the opposite direction and align full ESTs to the genome as done at Ensembl/UCSC and use these sequences to confirm introns. Comparing the results of Ensembl's alignment of ESTs to chromosome 22 with the results reported here would be an interesting project, but time to complete this work was not available.