The uniqueome: a mappability resource for short-tag sequencing

Ryan Koehler\(^1\), Hadar Issac\(^2\), Nicole Cloonan\(^3\) and Sean M. Grimmond\(^3\)

\(^1\)VerdAscend Sciences, West Linn, OR 97068, \(^2\)Imagenix Technologies, 2672 Bayshore Parkway Suite 502, Mountain View, CA 94043, USA and \(^3\)Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, St Lucia 4072, Australia

ABSTRACT

Summary: Quantification applications of short-tag sequencing data (such as CNVseq and RNAseq) depend on knowing the uniqueness of specific genomic regions at a given threshold of error. Here, we present the ‘uniqueome’, a genomic resource for understanding the uniquely mappable proportion of genomic sequences. Pre-computed data are available for human, mouse, fly and worm genomes in both color-space and nucleotide-space, and we demonstrate the utility of this resource as applied to the quantification of RNAseq data.


Supplementary information: Supplementary data are available at Bioinformatics online.

Received on April 12, 2010; revised on October 6, 2010; accepted on October 29, 2010

Massively parallel short-tag (25–100 nt) sequencing technologies are enabling a large repertoire of genomic and genetic research due to the depth of coverage that can be achieved in a cost-effective manner. Although short tags are most informative if they can be aligned uniquely to a reference genome, repetitive elements are not randomly distributed throughout the genome (Campbell et al., 2008); therefore, the proportion and location of uniquely mappable short sequences will also be non-randomly distributed. This presents a specific problem where quantitative comparison between two or more genomic regions is required (such as RNAseq or CNVseq).

For any quantitative analysis, it is desirable to understand the boundaries of the unique genome (the uniqueome), so that the amount of uniquely mappable sequence can be used to normalize tag counts. Uniqueomes have been studied comprehensively for small genomes with both long (Chaisson et al., 2004) and short (Whiteford et al., 2005) sequencing tags. For mammalian genomes, where comprehensive studies can be computationally prohibitive, the problem has been tackled with simulation (Campbell et al., 2008), region-specific computation (Robertson et al., 2008) or computation without mismatches (Rozowsky et al., 2009). Counterintuitively, considering only tags that align uniquely without mismatches does not resolve the problem of ambiguous mapping. In cases where the error rate of the sequencing platform exceeds the number of mismatches allowed during alignment, false positive uniquely aligning tags will occur (Supplementary Figure S1). It is therefore important to compute the uniqueome allowing for at least the number of errors likely to be present in the data.

We have used the exhaustive alignment feature of ISAS (Imagenix, USA) to systematically generate uniqueome data for human (hg18 and hg19), mouse (mm9), worm (ce6) and fly (dm3) genomes in both color-space and nucleotide-space. Ungapped alignments were performed independently for tag lengths between 25 and 90 nt with varying numbers of mismatches, in both nucleotide-space and color-space (Supplementary Material).

To visualize the results, non-unique genomic regions are formatted as bigBED and bigWig files, and these can be loaded directly into the UCSC genome browser (Kuhn et al., 2009). The BED files are also compatible with large-scale genomic analysis using the Galaxy interface (Goecks et al., 2010). Figure 1 illustrates the utility of uniqueome in identifying problematic alignment areas in an RNAseq dataset (Guttman et al., 2010).

Table 1 and Supplementary Tables S1–S4 describe the proportion of unique start sites and unique coverage for different genomes and different tag lengths in both nucleotide-space and color-space. Interestingly, increasing the length of the tag beyond 50 bp does little to overcome redundancy issues in mammalian genomes, suggesting that short-read technologies do not need to progress significantly beyond their current lengths to achieve optimum utility in fragment datasets.

\( ^* \)To whom correspondence should be addressed.

© The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Table 1. Proportions of unique start sites for nucleotide-space short tag alignments

<table>
<thead>
<tr>
<th>Species</th>
<th>25 (%)</th>
<th>30 (%)</th>
<th>35 (%)</th>
<th>50 (%)</th>
<th>60 (%)</th>
<th>75 (%)</th>
<th>90 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>66.0</td>
<td>70.9</td>
<td>74.1</td>
<td>76.9</td>
<td>77.5</td>
<td>79.3</td>
<td>80.8</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>69.9</td>
<td>74.4</td>
<td>77.1</td>
<td>79.1</td>
<td>79.4</td>
<td>80.7</td>
<td>81.7</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>85.3</td>
<td>87.7</td>
<td>89.0</td>
<td>89.8</td>
<td>89.9</td>
<td>90.6</td>
<td>91.1</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>67.5</td>
<td>68.4</td>
<td>69.0</td>
<td>69.2</td>
<td>69.2</td>
<td>69.5</td>
<td>69.8</td>
</tr>
</tbody>
</table>

Columns shown are length of tag matched; numbers in parentheses represent the number of mismatches allowed.

To better understand the effect of mapping uniqueness on RNAseq quantification, we determined the proportion of uniquely mappable positions in the RefSeq set of genes (Pruitt et al., 2007) using hg19 coordinates were investigated for 50mer tags using two mismatches in nucleotide-space and five mismatches in color-space.

Fig. 2. A mirror image plot showing the relationship between the length of a gene and the unique length of a gene for color-space (red) and nucleotide-space (blue). The uniqueness of human RefSeq genes (release 39) using hg19 coordinates were investigated for 50mer tags using two mismatches in nucleotide-space and five mismatches in color-space.

Table 2. Strategies to deal with multimapping tags and their correlation to microarray data from the same RNA sample

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw tag counts (RPKM)</td>
<td>0.38</td>
<td>0.35-0.41</td>
</tr>
<tr>
<td>Non-unique tag rescue counts (RPKM)</td>
<td>0.46</td>
<td>0.43-0.49</td>
</tr>
<tr>
<td>Uniqueome normalized tag counts (RPKM)</td>
<td>0.50</td>
<td>0.47-0.52</td>
</tr>
</tbody>
</table>

ACKNOWLEDGEMENTS

We thank D.F. Taylor, D.T.P. Tang, S. Wood and J.V. Pearson for server support; N. Waddell, K. Kassahn and the RGASP2 members for their helpful discussions.

REFERENCES


