BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 2 2011, pages 272–274
doi:10.1093/bioinformatics/btq640

Genome analysis

Advance Access publication November 12, 2010

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

Associate Editor: John Quackenbush

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 nucletotide-space, and we demonstrate the utility of this resource as applied to the quantification of RNAseq data.

Availability: Files, scripts and supplementary data are available from http://grimmond.imb.uq.edu.au/uniqueome/; the ISAS uniqueome aligner is freely available from http://www.imagenix.com/.

Contact: n.cloonan@uq.edu.au

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.

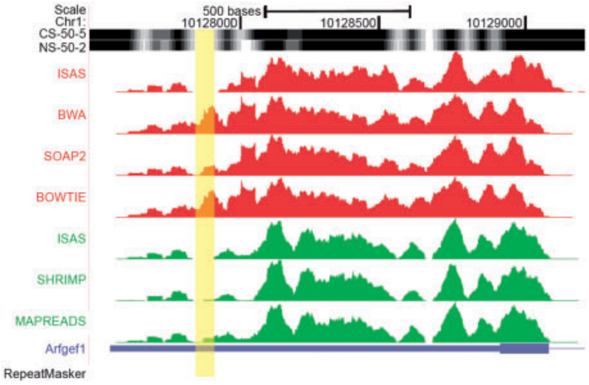

Fig. 1. Color-space (CS-50-5) and nucleotide-space (NS-50-2) uniqueome plots visualized alongside RNAseq data. The same 50mer RNAseq tags were aligned using several specialized short-read aligners in both nucleotide-space (red) and color-space (green). The yellow region highlights an area with no uniqueome coverage (confirmed by BLAT as a multimapping region), where tags have been falsely declared as 'uniquely mapping' by the various aligners. No repetitive elements were detected by RepeatMasker. See Supplementary Material for details.

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.

Table 1. Proportions of unique start sites for nucleotide-space short tag alignments

Species	25 (1) (%)	30 (1) (%)	35 (1) (%)	50 (2) (%)	60 (3) (%)	75 (4) (%)	90 (5) (%)
Homo sapiens[a]	66.0	70.9	74.1	76.9	77.5	79.3	80.8
Mus musculus[b]	69.9	74.4	77.1	79.1	79.4	80.7	81.7
Caenorhabditis elegans[c]	85.3	87.7	89.0	89.8	89.9	90.6	91.1
Drosophila melanogaster[d]	67.5	68.4	69.0	69.2	69.2	69.5	69.8

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

[a] Build hg19.

[b] Build mm9.

[c] Build ce6.

[d] Build dm3.

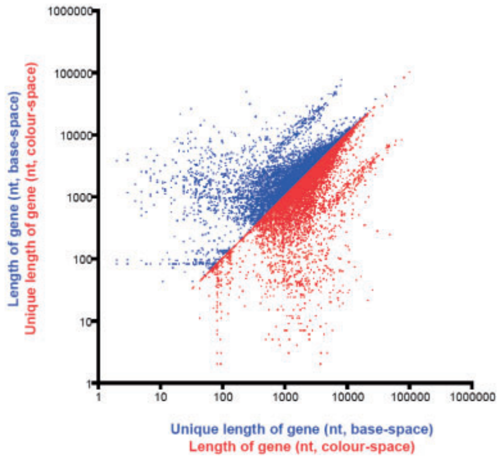

Fig. 2. A mirror image plot showing the relationship between the length of a gene and the unique length of a gene for color-space (red) and nucleotide-space (blue). The uniqueomes of human RefSeq genes (release 39) using hg19 coordinates were investigated for 50mer tags using two mismatches in nucleotide-space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

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) for 50mers in both color-space and nucleotide-space. Figure 2 shows a wide distribution of off-diagonal points reflecting the variability in the uniqueome content of RefSeq genes. Both the color-space and base-space plots reveal a group of RefSeq transcripts >5000 nt long but with less than 10% of uniquely mapping tags. This group of genes is highly enriched for large multicopy gene families, such as HLA. The uniqueness of RefSeq exon–exon junctions is described in Supplementary Tables S5 and S6.

Overall, the effect of non-unique short sequences in genes can be significant. More than 25% of RefSeq genes contain at least 10% of non-unique sequence when mapped as 50mers. Given that almost 40% of genes in mammalian genomes have arisen due to gene duplication (Zhang, 2003), this is not a surprising result. However, unless this is specifically normalized for in RNAseq experiments, this could bias both differential expression and gene set enrichment analyses. We have examined the utility of normalization using the uniqueome and compared it to both raw tag counts and non-unique tag rescue, using previously published sequencing and microarray data from the same samples (Cloonan *et al.*, 2008). Table 2 shows an improvement in the correlation of RNAseq to array data when using tag counts normalized to the proportion of unique sequence in each gene. Although the correlation improvements are lower than using a rescue approach, there is no additional computational time required to achieve this improvement, whereas significant CPU time is required for rescue (6 CPU hours using RNA-MATEv1.1; see Supplementary Material).

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

Method	Pearson	95% confidence interval
Raw tag counts (RPKM)	0.38	0.35–0.41
Non-unique tag rescue counts (RPKM)	0.46	0.43–0.49
Uniqueome normalized tag counts (RPKM)	0.50	0.47–0.52

Finally, the uniqueome allows higher confidence in mutation detection (e.g. cancer resequencing), where mis-mapping can confound SNP calling algorithms. This is a particular problem faced by users of paired-end or mate-pair data, where the mapping position of a multimapping tag is rescued based on its pair which uniquely maps. It is important to note that while this rescue can lead to improved levels of coverage (Bainbridge *et al.*, 2010), it does not increase the uniquely mapping proportion of the genome, and can lead to the misplacement of tags and false positive variant calls (Supplementary Figure S2). The uniqueome can be used to identify these regions of low confidence, independently of the aligner used to generate the data, as illustrated in Figure 1.

Although described here as a resource for short-tag sequencing applications, the utility of this resource extends beyond this theme. Primer design, comparative genomics and microarray probe design would also derive benefit from this resource. A PDF tutorial on using the uniqueome with Galaxy is provided (Supplementary Material). The ISAS uniqueome aligner is freely available, and a PDF tutorial on its use is provided (Supplementary Material).

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.

Funding: National Health and Medical Research Council (455857 to S.M.G. 456140, 631701); Australian Research Council (DP1093164 to N.C., DP0988754) in part.

Conflict of Interest: H.I. derives income from ISAS sales.

REFERENCES

Bainbridge,M.N. *et al.* (2010) Whole exome capture in solution with 3 Gbp of data. *Genome Biol.*, **11**, R62.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Chaisson,M. *et al.* (2004) Fragment assembly with short reads. *Bioinformatics*, **20**, 2067–2074.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Method*, **5**, 613–619.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.

Kuhn,R.M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Robertson,A.G. *et al.* (2008) Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res.*, **18**, 1906–1917.

Rozowsky,J. *et al.* (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

Whiteford,N. *et al.* (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.*, **33**, e171.

Zhang,J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.*, **18**, 292–298.