Accessing the Inaccessible Genome

Designing mate pair libraries to distinguish disease loci within large duplications

Heather E. Peckham[1], Timothy T. Harkins[2] and Ellen M. Beasley[2], Life Technologies, [1]Beverly, MA, USA, [2]Foster City, CA, USA

ABSTRACT

Spinal muscular atrophy (SMA) is a neuromuscular disease that results in progressive muscular weakness. While it is the leading cause of infant death, it is considered an orphan disease that receives relatively little research funding due to the small patient population. SMA is caused by mutations in the SMN1 (Survival Motor Neuron) gene. Humans have a second gene SMN2 which differs from SMN1 by a single point mutation and produces mostly nonfunctional protein. When there is no functional SMN1 gene, disease severity depends primarily on the number of SMN2 gene copies. SMN1 and SMN2 fall within a 500 kb inverted duplication with repetitive elements that make it prone to rearrangements and deletions. Complex regions containing duplicate copies of genes pose a challenge to next-generation sequencing efforts since these regions are generally not mappable and considered inaccessible. While next-generation sequencing has opened doors to many orphan diseases, SMA represents a disease in which these technologies struggle without a careful and deliberate approach to re-sequencing. Inaccessible regions such as these have the potential to be accessed with strategies involving use of a reference sequence to uncover unique loci within the duplications and the design of libraries with appropriately sized molecules such that they overlap each other and reach into non-unique regions. The reference sequence can also be used to elucidate mapping strategies that allow reads to be unambiguously aligned to regions containing only slight molecular differences. These approaches enable the 'unmappable' SMN genes to be distinguished from each other as well as phased into separate haplotype structures. We use next-generation sequencing of whole human genomes to elucidate the complex properties of the SMN genes as an example of how to design sequencing strategies, including selecting library types, insert sizes, read lengths and analysis procedures, that are necessary to reveal complex genomic features.

INTRODUCTION

The genetic structure of SMN1 and SMN2 is unique to humans. The two genes fall within a 500 kb inverted duplication which contains telomeric (SMN1) and centromeric (SMN2) copies (Figure 1). There are 5 nucleotide differences between SMN1 and SMN2 of which 2 are exonic. One exonic difference is in the 3' untranslated region of exon 8 and the other is a translationally silent change of C->T in exon 7. The new nucleotide sequence of exon 7 of SMN2 eliminates the motif that is recognized by a splicing factor and thus prevents the splicing of the exon. Approximately 10% of the resulting SMN2 transcripts contain full-length SMN protein and in the absence of functional SMN1, the severity of SMA depends on the number of copies of SMN2 that are present in genomic DNA.

Mutations in SMN1 are known to exist in 3 forms:

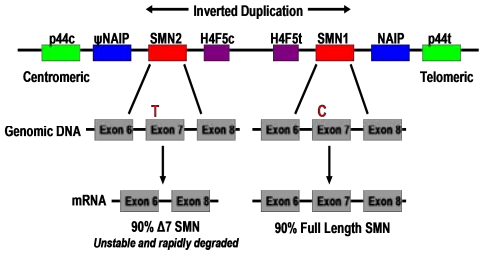
1) Deletions that cause partial or complete removal of the SMN1 gene

2) Conversion in which SMN1 is converted into an SMN2-like gene due to a C->T change in exon 7

3) Point mutations that result in the production of non-functional or unstable SMN protein

A comprehensive approach to determining the SMN status of an individual includes determining the existence, structure and number of copies of SMN1, SMN2 and SMN2-like genes. A strategy to successfully align reads to the C/T transition of exon 7 in SMN1 and SMN2 in a given individual begins before sequencing and includes determining the mappability of the genome with a given read length, finding loci that are unique in the vicinity of SMN1 and SMN2 and designing a mate pair library that will allow one tag to be uniquely placed and the other tag to reach in to the C/T transition of exon 7 and be unambiguously placed in either SMN1 or SMN2.

Figure 1. The SMN1 and SMN2 genes fall within a 500 kb inverted duplication.

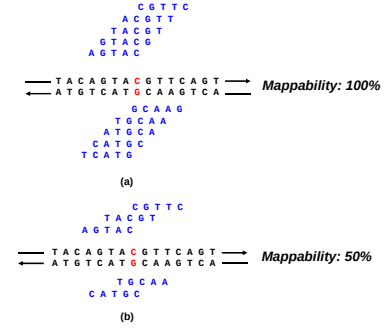


A translationally silent change of C->T in exon 7 of SMN2 prevents the splicing of the exon. Approximately 10% of the resulting SMN2 transcripts contain full-length SMN protein. In the absence of functional SMN1, the severity of SMA depends on the number of copies of SMN2 that are present in genomic DNA.

METHODS

- Calculate the mappability of the genome (Figure 2)
  - Determine the loci within the region of interest that are uniquely mappable within the entire genome. The in-silico mappability of a reference sequence is determined by taking each n-mer (n is the tag length) in the reference and attempting to map it back to the genome. When using 50 bp reads, the mappability of any locus will range from 0 to 100 since there are up to 50 50-mers on each strand that may be able to map uniquely back to the place in the genome from which they were derived.
- Find loci that are unique between the SMN1 and SMN2 regions (Figure 3)
  - Compare the regions upstream and downstream of exon 7 within the SMN1 and SMN2 duplications to find loci that are unique to each duplication
    - Assess the mappability of each locus
    - Assess whether each locus is in a group of unique loci
- Select unique loci with the highest mappability and largest group size
  - Determine the distance between the selected loci and exon 7
  - Construct a mate pair library consisting of molecules the size of the distance between the selected loci and exon 7 so that they can anchor at the unique loci and reach into the C/T transition in exon 7

Figure 2. Each locus in a genome has a degree of mappability between 0% and 100%.



The mappability of a genome is calculated by matching a uniformly fragmented genome to itself. The mappability of a given locus using 5 bp reads is illustrated. Examples of 100% and 50% mappability are shown in (a) and (b), respectively.

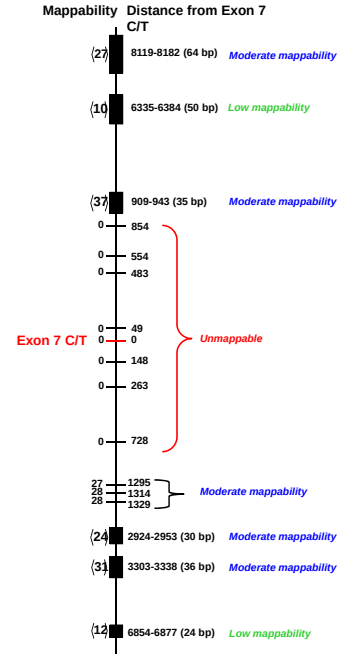
RESULTS

- The C/T transition in exon 7 of SMN1 and SMN2 is not uniquely mappable with 50 bp reads. A mate pair library can be used to align the reads originating from these regions in which one tag is placed in a unique region outside of these sites and the other tag is able to reach in and unambiguously place at the C/T transition in exon 7.
- The unique loci closest to exon 7 are unmappable in the context of the entire genome and thus molecules larger than these distances must be selected.
- Three unique loci ~1300 bp from exon 7 have a moderate amount of mappability but the existence of only 3 unique bp in this vicinity prevents it from being an ideal choice.
- Stretches of DNA 6.3 kb and 6.8 kb from exon 7 have unique loci but low mappability and thus most molecules originating from these areas would not be able to uniquely place the anchor tag.
- Molecule sizes in which an anchor tag could be placed within a stretch of unique DNA with moderate mappability and another tag could be placed unambiguously in exon 7 include:
  - 900 bp: anchor within a 35 bp region that is on average 37% mappable
  - 2.9 kb: anchor within a 30 bp region that is on average 24% mappable
  - 3.3 kb: anchor within a 36 bp region that is on average 31% mappable
  - 8 kb: anchor within a 64 bp region that is on average 27% mappable

CONCLUSIONS

- Disease loci may reside within large duplications that are considered to be unmappable
- Mate pair libraries have the ability to anchor one tag within unique sequence and reach into unmappable regions with the other tag
- Large duplicated regions can be examined prior to sequencing in order to design mate pair libraries of the appropriate size so that they can reach from loci that are both unique between the duplications and uniquely mappable in the entire genome and into the region of interest

Figure 3. Molecule sizes can be selected such that they can anchor in a unique region and extend into an inaccessible region.



A mate pair library can be selected in which one tag of a given molecule will map to a locus that is both unique between the SMN1 and SMN2 regions and uniquely mappable in the genome and the other tag will reach into the C/T transition in exon 7.

ACKNOWLEDGEMENTS

We thank Yutao Fu for providing mappability data and insight.

REFERENCES

1. The genetics of SMA - http://www.fsma.org/
2. Motor neuron disorders and related diseases, Vol 82, Andrew Eisen and Pamela J. Shaw.
3. http://neuromuscular.wustl.edu/synmot.html#neighbor