1. Important Terms

This manual is intended to be used by geneticists. It assumes that the reader has basic understanding in molecular biology and genetics. The following terms are crucial in understanding the material in this manual:

- **Genetic Markers**: A short stretch of DNA that can be identified by various techniques in the laboratory. These are used to detect a particular gene that may be close to the marker along the DNA.

- **Genetic Mapping**: The use of genetic markers to map important mutations and genes along the chromosomes of a specific organism.

- **Genetic Mapping in Mice**: Two different strains of mice are crossed, one containing the mutation of interest. The resulting offspring are tested for whether they contain the genetic markers that are potentially close to the mutation of interest along the DNA. As this proceeds, the region containing the mutation is narrowed down to a particular location.

- **SNP**: Single Nucleotide Polymorphism. There are different nucleotide permutations at a single position in the DNA in different individuals in the population. For example, the general population may have an A nucleotide at a certain position, but some individuals have a C, T or a G nucleotide at that position. SNPs are loosely referred to as “point mutations”.

- **SNP Markers**: These polymorphic sites can be used to fine map a region containing a gene or mutation of interest. Often the exact location of a gene or mutation is unknown, but the locations of SNPs possibly surrounding the gene are known. By using simple techniques in the laboratory, these SNPs can be detected in an individual.

- **Restriction Enzyme**: A protein that cuts DNA at particular recognition sites. For example, the enzyme EcoRI cuts at any part of a DNA strand that contains the nucleotide pattern GAATTC.

- **RFLP**: Restriction Fragment Length Polymorphism. A strand of DNA is digested with a particular restriction enzyme that cuts at a particular site. If the recognition site for that enzyme is disrupted by a SNP then the enzyme will not cut. An RFLP assay is a simple way to detect whether an
individual has the SNP marker of interest. If the person does have the SNP then there will be one long strand of DNA in the assay. If the person does not have the SNP then the enzyme will cut and there will be two smaller strands of DNA shown in the assay.

- **Informative Marker**: In the case with SNPs, an informative marker is one that will produce a valuable RFLP that can be assayed by using DNA from the offspring of the mice that are initially crossed in the genetic mapping experiment.

- **PCR**: Polymerase Chain Reaction. A laboratory method that can be used to amplify a specific region of DNA into millions of copies. It requires that two primers, one on each end of the region of DNA to be amplified, in order for the chain reaction to start.

- **Phenotype**: An observable characteristic in an organism that is dictated by genetic and environmental factors.

- **Genotype**: The genetic makeup of an organism at a particular location in the genome.

## 2. Introduction

Many mouse gene mutations cause phenotypes that serve as models of human genetic disorders. By mapping a mutation onto the mouse chromosomes we can accelerate our understanding of the mouse gene that contains the mutation, its human counterpart gene, and the underlying etiology of the disorder caused by the mutation. SNP markers have been shown to give high resolution in genetic mapping because they are abundant throughout the genome and can be analyzed in a high-throughput manner using automated technology [1]. A database of known SNPs (dbSNP) has been constructed that contains the genotypes for SNPs in different strains of mice [2,3,4]. Recently, SNP markers have been used in genome-wide SNP panels to simplify genetic mapping in mouse [5]. However, initial localization via a SNP panel often defines a large chromosomal interval where a mutation of interest may lie, and “bench-top” technologies for fine-mapping using SNPs are often inefficient. We have developed a web-based tool, SNP2RFLP, that can extract region-specific SNPs from dbSNP and identify those SNPs that create restriction fragment length polymorphisms (RFLPs) that can be
easily assayed by digesting SNP-containing DNA with specific restriction enzymes after it has been amplified by PCR.

The input to SNP2RFLP is the two mouse strains used in the cross, the chromosomal region, and a set of restriction enzymes. SNP2RFLP extracts the SNPs from dbSNP that are polymorphic between the two strains in the region in question. The program simulates a restriction digest of the SNP-containing sequences with each enzyme to determine whether the SNP creates an RFLP. Informative markers are then analyzed using the Primer3 program [6], which finds suitable PCR primers surrounding the SNP. The output of SNP2RFLP is the informative SNPs that create RFLPs and the left and right PCR primers. This information can then be used to readily perform the RFLP assays and further refine the region containing the mutation of interest.

This manual goes into the details of the required input to SNP2RFLP, the process by which informative markers are found, and the output to SNP2RFLP. SNP2RFLP was developed by the Sunyaev and Beier Laboratories in the Division of Genetics at Brigham and Women’s Hospital, in the Harvard Medical School (http://genetics.bwh.harvard.edu/genetics/labs.html). The tool was primarily implemented by Wes Beckstead and comments and questions about the use and operations of SNP2RFLP should be directed to him by email at wesb@byu.edu.

3. Input to SNP2RFLP

The following is a detailed description of the input needed for SNP2RFLP to perform its analysis.

3.1 The Two Mouse Strains Used in the Cross.

SNP2RFLP checks for informative SNPs between the two mouse strains that are being crossed in the genetic mapping experiment. The local database that SNP2RFLP uses contains the genotypes for SNPs in 99 different mouse strains such as FVB or A/J. The user interface to enter the two mouse strains is shown below.
The majority of the 99 strains used by the program have only hundreds or thousands of known genotypes for SNPs in the database. Using these strains does not often return any helpful results strictly due to the lack of information on these strains in the database. The following table shows the most useful strains to be used in SNP2RFLP. These have on the order of millions of genotypes for known SNPs in the database.

SNP2RFLP will not proceed with its analysis if the mouse strain the user chooses for strain 1 is the same as the choice for strain 2.
SNP2RFLP uses a list of commonly used restriction enzymes to perform its analysis. These are as follows:

- AluI
- EcoRV
- HinfI
- MspI
- RsaI
- ScrFI
- AflII
- Fnu4HI
- KpnI
- PstI
- SacI
- Sau96I
- ClaI
- HaeIII
- MboI
- PvuI
- SalI
- Sau96I
- DdeI
- Hhal
- MseI
- PvuII
- Scal

These enzymes are used by default to simulate the restriction digest described in the introduction, but additional enzymes can be selected from a list of over 1,300 restriction enzymes. Additionally, the “select all enzymes” option can be selected to include all the enzymes in the simulated restriction digest. All the restriction enzymes and recognition sequences used by SNP2RFLP were obtained from the restriction enzyme database (REBASE) [7,8]. The user interface to enter the restriction enzymes is shown below.

Caution should be taken when using the “select all” option. If used in conjunction with a large genomic region (see section 3.3), then the results may take some time to load.
As discussed in section 2, genetics mapping experiments using a genome-wide SNP panel defines a large genomic region where the mutation of interest may be located. SNP2RFLP requires that this region is given in order for informative SNP markers in this region to be found. These SNP markers can then be used to fine map this initial genomic region and locate the mutation causing a disorder. The program requires the chromosome number and the starting and ending base or nucleotide numbers (e.g. chromosome 1, begin: 1000024, end: 1000100) as shown below.

![Chromosome: 9]

**Base Position: From** [ ] **To** [ ]

SNP2RFLP will not proceed unless the base positions are valid nucleotide positions on the chromosome chosen. Also, the beginning base number should be less than the ending base number.

**WARNING**
A large genomic region (greater than 100000000 bases) may take a long time to load in a browser.

### 3.4 Options to Control the Output

In many cases, the number of informative SNPs returned is more than needed, or in other cases the number of SNPs returned is not adequate. There are multiple options to control the output returned by SNP2RFLP to produce a desired number of informative markers. First, each SNP in the database has a validation status. NCBI’s dbSNP defines many different ways that a SNP can be validated. For the sake of simplicity, if the “display validated SNPs only” option is selected, SNPs
that have no validation information at all are excluded. This reduces the number of informative SNPs in many cases, but gives higher confidence in the results.

Second, there are many times when no informative SNPs can be found between two strains in a specific region. It may be the case that there are many SNPs in the region but the genotype may only be recorded in one strain. If the “display SNPs recorded in only one strain” option is selected then the restriction digest is simulated on the SNPs from the strain where the genotype is known and additional methods can be used by the user to infer the genotype of each SNP in the other strain.

Third, amplifying the region around a SNP with PCR is often difficult if the SNP is found in a repeat region of the genome. The user can select the “disregard SNPs found in repeat regions” option and the program will not display these SNPs. SNP2RFLP tests whether a SNP falls in a repeat region using the mouse genome premasked by RepeatMasker [9]. This will often reduce the output and guarantee that difficult SNPs to amplify will not be included.

Fourth, the SNP2RFLP output page is very instructive (see section 5). But it is often useful for many who use the tool to have the output come in the form of a tab-delimited text file so that the informative SNPs can be viewed in a spreadsheet of some sort. If the “produce a tab-delimited text file” option is selected then a window will pop up along with the normal SNP2RFLP results page. This file can then be saved using the browser’s File -> Save Page As menu.

Lastly, the desired density of SNP markers returned can be set. SNP2RFLP can be told to return all of the informative markers, 1 every 5, 1 every 20 etc. This is an extremely valuable option which allows the user to retrieve an adequate amount of output that is also manageable. These output controls are shown below.
With the required input, SNP2RFLP performs its analysis in the following steps:
1) extract SNPs and flanking sequences from the database, 2) simulate restriction digest with the selected enzymes, and 3) design PCR primers with Primer3. These three steps are described below.

**Output Control**
- Display validated SNPs only
- Display SNPs that are only recorded in one strain but may be verified in the other
- Don't display SNPs that are found in repeat regions
- Produce a tab-delimited test file of the results

Density of SNPs (all, 1 every 5, 1 every 20, etc.)

4. SNP2RFLP Analysis

With the required input, SNP2RFLP performs its analysis in the following steps: 1) extract SNPs and flanking sequences from the database, 2) simulate restriction digest with the selected enzymes, and 3) design PCR primers with Primer3. These three steps are described below.

4.1 Extract SNPs and Flanking Sequences

With the required input SNP2RFLP retrieves all of the SNPs and their flanking sequences from the local database that are polymorphic between the two strains in the genomic region in question. Shown below is a SNP and the flanking sequences extracted from the database. The SNP is enclosed in brackets and highlighted in red. The two permutations of the single nucleotide at that position are separated by a forward slash. For example, in the sequence below, if the strains FVB and B6 were being crossed in the genetic mapping experiment A would be the genotype of FVB at that position and C would be the genotype of B6 at that same site.

**CACACAGAGAT**

**CAAGCGAGGG**

Discard: Non-Informative

SNPs that happen to show the same genotype in both strains are discarded as they will not produce meaningful RFLP assays. SNP2RFLP is only interested in SNPs that are polymorphic between the two strains being crossed in the genetic mapping experiment.

**GGAACGTGTGT[A/C]CCAACAGTTAC**

**CACACAGAGAT[C/C]CAAGCGAGGG**
SNP2RFLP then simulates a restriction digest on the SNP-containing sequences with each enzyme that was selected (see section 3.2) to see if the SNP creates an RFLP. In other words, a SNP-containing sequence is scanned to see if the recognition sequence for the particular enzyme is found in strain A but not found in strain B due to the alteration of the recognition sequence by the SNP. If this situation occurs, the SNP is considered informative as it can be easily assayed by amplifying the region with PCR, digesting the products with the enzyme, and examining the lengths of the resulting DNA strands using simple techniques in the laboratory. Shown below is an example of an informative SNP. The restriction enzyme that is being tested has a recognition sequence of TGTA. The SNP between strain 1 and strain 2 allows this enzyme to cut in strain 1 but not in strain 2, making an RFLP that can be assayed at the benchtop.

**Enzyme Recognition Sequence: TGTA**

**Informative**

Strain 1: GGAACGTG\textbf{TGT}\textit{A}CCAACAGTTAC

Strain 2: GGAACGTG\textit{TGT}CCAACAGTTAC

The example shown below is not an informative SNP because the simulated digest shows that the SNP will not create an RFLP. For the enzyme being tested, the recognition site is not found in either of the SNP-containing sequences.

**Enzyme Recognition Sequence: ACTG**

**Discard, Non-Informative**

Strain 1: GGAACGTG\textbf{TGT}CCAACAGTTAC

Strain 2: GGAACGTG\textit{TGT}CCAACAGTTAC

And the example below shows a SNP in which the enzyme being tested cuts in both strains because the recognition site is not found at the SNP position. This SNP would not be informative because there is no way to differentiate between the strains after the digest is performed and the region amplified. There would be the same lengths of DNA strands regardless of which strain the DNA came from.

**Enzyme Recognition Sequence: AACA**

**Discard, Non-Informative**

Strain 1: GGAACGTG\textbf{TGT}ACC\textbf{AACA}GTTAC

Strain 2: GGAACGTG\textit{TGT}CCC\textit{AACA}GTTAC
4.3 Design PCR primers with Primer3

SNP2RFLP not only tells the user what SNPs will be informative between the two strains being crossed in the genomic region in question, it also goes ahead and designs the PCR primers that the user will use to amplify the regions containing the informative SNPs. Each SNP-containing sequence is analyzed with the Primer3 program which designs appropriate left and right PCR primers surrounding the SNP. An example of a SNP-containing sequence is shown below with the left and right primers highlighted. The left primer matches the nucleotide sequence so it can anneal to the DNA strand shown. The right primer is the reverse complement so it can anneal to the strand opposite the DNA shown.

GGAACGTGTGT[A/C]CCAACAGTTAC

GAACG       AACTG

5. SNP2RFLP Output

After SNP2RFLP performs its analysis the output is loaded on a separate web page. The output is the informative SNP markers ordered by position in the genome, the sites at which the selected enzymes cut, and the appropriate left and right primers mapped onto the sequence. Shown below is an example of a single SNP marker that is returned by SNP2RFLP.

<table>
<thead>
<tr>
<th>Position</th>
<th>SNP ID</th>
<th>A/J</th>
<th>FVB/NJ</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>26837654</td>
<td>29038413</td>
<td>A</td>
<td>C</td>
<td>GCTCAGTGCCCACAATGCTAGAT[A/C]CTATGGCAATTCAGCTAGTGAAG</td>
</tr>
</tbody>
</table>

The position is linked to the NCBI map viewer showing the region in the genome where the SNP lies. The SNP ID is linked to the reference SNP ID page also on the NCBI website. The genotypes for each strain are also linked to the NCBI
6. Output Control

The ability to control the amount of information returned by SNP2RFLP is important. As an example, suppose a genome-wide SNP scan, crossing A/J and FVB, reveals a candidate region on chromosome 13, nucleotide 14800000 to nucleotide 46700000, in which a particular mutation of interest may be located. SNP2RFLP may be used to find additional SNP markers to fine-map this region and reveal the exact location of the mutation. The table below shows the different numbers of informative SNPs in this region returned by SNP2RFLP by selecting different options in the program. When selecting all of the enzymes and telling SNP2RFLP to keep all SNPs, a large and perhaps unmanageable number of SNP markers is returned. By restricting the enzymes and telling SNP2RFLP the desired density of SNPs, a manageable and adequate number of SNP markers can be displayed. This will facilitate fine-mapping of this region and locating the mutation causing the phenotype of interest.

<table>
<thead>
<tr>
<th>All Enzymes, Keep all SNPs</th>
<th>Default Enzymes, Keep all SNPs</th>
<th>Default Enzymes, Keep 1 every 5 SNPs</th>
<th>1 Particular Enzyme, Keep all SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>935 SNP Markers</td>
<td>257 SNP Markers</td>
<td>52 SNP Markers</td>
<td>39 SNP Markers</td>
</tr>
</tbody>
</table>

7. Conclusion

Mapping mouse gene mutations onto their chromosomal locations is valuable for understanding human disorders caused by similar mutations. Initial localization of a mutation causing a phenotype of interest defines a large genomic region and benchtop technologies with which to proceed with fine-mapping are limited. We have utilized a database of SNPs (dbSNP) and a database of restriction enzymes (REBASE) to create a web-based tool, SNP2RFLP, that facilitates genetic mapping in mouse by extracting additional SNP markers that can be used to
narrow this initial interval and find the exact location of the mutation in the mouse genome. SNP2RFLP offers options to control the number of SNP markers returned which makes for adequate and manageable output available to the user. SNP2RFLP is freely available and should prove very useful to the general mouse genetics community.

# 8. Appendix and References

## 8.1 Troubleshooting Guide

<table>
<thead>
<tr>
<th>Problem</th>
<th>Potential Causes</th>
<th>Solutions</th>
</tr>
</thead>
</table>
| Error messages appear on the screen and SNP2RFLP won’t proceed to the results page. | SNP2RFLP will not perform its analysis unless the program has validated that all input fields are correct. | 1. Make sure you have entered both mouse strains, and make sure that they are different from each other.  
2. Make sure you have entered both beginning and ending base numbers. The base numbers must be valid for that chromosome (e.g. a base number of -1 or 10000000000 does not exist on chromosome 1 in mice) and make sure that your beginning base number is less than your ending base number.  
3. Make sure you enter the density field on the output control correctly. Use “all” if you want SNP2RFLP to keep all SNPs found, or use “1 every 5” or “1 every 3” for example to keep every 5th and 3rd SNP respectively. |
| I can’t select multiple enzymes.                                        | You may not be using the control or command keys correctly.                      | To select multiple enzymes you must hold down the control key if you are using a PC or hold down the command key if you are using a MAC while you are clicking on multiple enzymes. Alternately, you can simply use the “select all enzymes” option but this may produce more results than desired. |
| I get the warning message saying that the results may take a long time to load.    | The genomic region you selected is too large.                                    | SNP2RFLP’s local database has millions of SNPs in it and if a large genomic region is chosen with two strains that each have millions of SNPs in that region then SNP2RFLP may take a long time to load. Try splitting the genomic region up and only run part of it at a time. |
| My results page takes too long to load all the way.                    | The genomic region you selected is too large.                                    | Use a smaller genomic region. Try splitting up the large region into multiple smaller regions and run them one at a time. |
| I don’t get a tab-delimited text file of the results.                  | 1. You haven’t selected the right output control options.                       | 1. Make sure you select the “tab-delimited text file” option on the output control.  
2. Your browser is blocking pop-ups from SNP2RFLP.                      | 2. Check your browser’s pop-up blocker and tell it to allow pop-ups from http://genetics.bwh.harvard.edu/snp2rflp. |
SNP2RFLP does not return any informative SNPs.

1. Your genomic region is too small.
2. Not enough restriction enzymes are included in the search.
3. You are using strains that do not have any SNPs in the selected region.

1. Choose a larger genomic region.
2. Include more restriction enzymes in your search or use the “select all enzymes” option.
3. Some strains in the database do not have very many SNPs to actually be tested by SNP2RFLP to see if they are informative. It is best to use the strains shown in the table in section 3.1 because they have on the order or millions of SNPs in the database.

8.2 References


