**Introduction**

Peptide and protein identification from tandem mass spectrometry data is usually performed by searching data against a protein database that is derived from translating the sequenced genome from a representative of the species into protein sequence. The weaknesses of this approach are:

1. A subset of genes are expressed as proteins in a given tissue under a certain set of conditions;
2. For a given protein there may be splice variants expressed that are not represented in any database;
3. Variants (e.g. due to single nucleotide polymorphisms) may not be in the protein database.

Sequencing technology has now become sufficiently fast and cheap that it is practical to sequence all RNA expression in a specific tissue or cell line under defined conditions. By constructing a reference database from this RNA-Seq data, it should be possible to identify peptides formed by novel gene splicing or sequence variants that are not included in a reference protein database.

The Proteome Informatics Research Group (iPRG) this year performed a study to evaluate the benefits of using databases derived from RNA-Seq data for peptide identification. The proteomic database provided consisted of high mass accuracies, a variety of human peripheral blood mononuclear cells. A diversity of different types of sequence databases were supplied. These included a standard protein sequence database, a database containing only sequences of proteins expressed in the sample based on RNA-Seq data, a database that included sequence and splice variants, a database of peptides that could not be recognized by human expressed gene sequences.

Participants were asked to report spectral identifications in the form of an Excel spreadsheet highlighting those identifications that were only identified using one of the RNA-Seq derived specialized sequence databases. Participants were also required to complete a web-based questionnaire summarizing the tools and methods they used.

**Study Goals**

1. Use a dataset with matched RNA-Seq and tandem mass spectrometry data.
2. By comparing RNA-Seq data to reference genome sequence create two extra databases.
3. Sequences corresponding to SNV kinase as compared to reference genome sequence.
4. Novel sequence that did not match reference genome called for a SNV.
5. Allow participants to use the bioinformatics tools and methods of their choice.
6. View a common reporting template.
7. Report results at an estimated 1% FDR (at the peptide level).
8. Ignore protein reference.

**Study Materials**

- "41LC MS/MS files
- RFLP-mcrf or MIG – conversions by mouse/rodent (Protectogen)
- RNA-Seq data
- Four reference protein databases derived from RNA-Seq data
- Results template (Excel) 
- On-line survey (Survey Monkey)

**Study Instructions**

1. Retrieve and analyze the data file in the format of your choosing, with the methods of your choosing.
2. Use a reference database and compare results from other databases to those identified in reference database. Report the peptide to spectrum matches in the provided template.
3. Fill out the survey.
4. Attach a 1-2 page description of the methodology employed.

**Introduction**

Peptide and protein identification from tandem mass spectrometry data is usually performed by searching data against a protein database that is derived from translating the sequenced genome from a representative of the species into protein sequence. The weaknesses of this approach are:

1. A subset of genes are expressed as proteins in a given tissue under a certain set of conditions;
2. For a given protein there may be splice variants expressed that are not represented in any database;
3. Variants (e.g. due to single nucleotide polymorphisms) may not be in the protein database.

Sequencing technology has now become sufficiently fast and cheap that it is practical to sequence all RNA expression in a specific tissue or cell line under defined conditions. By constructing a reference database from this RNA-Seq data, it should be possible to identify peptides formed by novel gene splicing or sequence variants that are not included in a reference protein database.

The Proteome Informatics Research Group (iPRG) this year performed a study to evaluate the benefits of using databases derived from RNA-Seq data for peptide identification. The proteomic database provided consisted of high mass accuracies, a variety of human peripheral blood mononuclear cells. A diversity of different types of sequence databases were supplied. These included a standard protein sequence database, a database containing only sequences of proteins expressed in the sample based on RNA-Seq data, a database that included sequence and splice variants, a database of peptides that could not be recognized by human expressed gene sequences.

Participants were asked to report spectral identifications in the form of an Excel spreadsheet highlighting those identifications that were only identified using one of the RNA-Seq derived specialized sequence databases. Participants were also required to complete a web-based questionnaire summarizing the tools and methods they used.

**Study Goals**

1. Use a dataset with matched RNA-Seq and tandem mass spectrometry data.
2. By comparing RNA-Seq data to reference genome sequence create two extra databases.
3. Sequences corresponding to SNV kinase as compared to reference genome sequence.
4. Novel sequence that did not match reference genome called for a SNV.
5. Allow participants to use the bioinformatics tools and methods of their choice.
6. View a common reporting template.
7. Report results at an estimated 1% FDR (at the peptide level).
8. Ignore protein reference.

**Study Materials**

- "41LC MS/MS files
- RFLP-mcrf or MIG – conversions by mouse/rodent (Protectogen)
- RNA-Seq data
- Four reference protein databases derived from RNA-Seq data
- Results template (Excel) 
- On-line survey (Survey Monkey)

**Study Instructions**

1. Retrieve and analyze the data file in the format of your choosing, with the methods of your choosing.
2. Use a reference database and compare results from other databases to those identified in reference database. Report the peptide to spectrum matches in the provided template.
3. Fill out the survey.
4. Attach a 1-2 page description of the methodology employed.