DataSet

- High level ‘container’ object
  - Fetched with DataSetAdaptor
- Links raw data to processed data
- Key attributes/methods:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Example value(s)</th>
<th>Method(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>RegulatoryFeatures:MultiCell</td>
<td>$ds-&gt;name</td>
</tr>
<tr>
<td>Product feature set</td>
<td>FeatureSet Object</td>
<td>$ds-&gt;product_FeatureSet</td>
</tr>
<tr>
<td>Supporting Sets</td>
<td>List of supporting Set objects e.g. ResultSet or FeatureSet</td>
<td>$ds-&gt;supporting_sets</td>
</tr>
</tbody>
</table>
FeatureSetAdaptor

- Provides fetch methods for FeatureSet objects:
- Key methods:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Example arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>fetch_by_name</td>
<td>‘cisRED motifs’</td>
</tr>
<tr>
<td>fetch_all_by_CellType</td>
<td>CellType object</td>
</tr>
<tr>
<td>fetch_all_by_FeatureType</td>
<td>FeatureType object</td>
</tr>
<tr>
<td>fetch_all_by_feature_class</td>
<td>external, annotated or regulatory</td>
</tr>
</tbody>
</table>
FeatureSet

- Processed Feature container:
  - Regulatory, Annotated, External, Segmentation
- Key attributes/methods:

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Example value(s)</th>
<th>Method(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>HepG2_USF1_ENCODE_Hudsonalpha_SWEMBL_R015</td>
<td>$fs-&gt;name</td>
</tr>
<tr>
<td>Display label</td>
<td>USF1 - HepG2 Enriched Sites</td>
<td>$fs-&gt;display_label</td>
</tr>
<tr>
<td>Cell Type (*)</td>
<td>K562</td>
<td>$fs-&gt;cell_type-&gt;name(*)</td>
</tr>
<tr>
<td>Feature Type</td>
<td>CTCF</td>
<td>$fs-&gt;feature_type-&gt;name</td>
</tr>
</tbody>
</table>
| Features              | List of Features e.g. Annotated or RegulatoryFeatures | $fs->get_Features_by_Slice
                                        |                                                   | $fs->get_all_Features                        |

* : Some Feature Sets may not have metadata like cell type
my $fset_adaptor =
    $reg->get_adaptor('Human', 'funcgen', 'featureset');

my @tfs =
    @{$ftype_adaptor->fetch_all_by_class('Transcription Factor')};

foreach my $ft (@tfs){
    my @fsets = @{$fset_adaptor->fetch_all_by_FeatureType($ft)};
    print "Found ".scalar(@fsets).' '.$ft->name." FeatureSets:

    foreach my $fset (@fsets){
        print "\t".$fset->name." - ";
        print "\t".$fset->display_label."\n";
        print "\t\t".$fset->cell_type->name;
        print "\t".$fset->feature_type->name;
        print "\t".scalar(@{$fset->get_all_Features})." AnnotatedFeatures
    }
}
my $fset_adaptor = $reg->get_adaptor('Human', 'funcgen', 'featureset');

my @tfs = @{$ftype_adaptor->fetch_all_by_class('Transcription Factor')};

foreach my $ft (@tfs) {
    my @fsets = @{$fset_adaptor->fetch_all_by_FeatureType($ft)};
    print "Found \t" . scalar(@fsets) . ' ' . $ft->name . ' FeatureSets:
';
    foreach my $fset (@fsets) {
        print '	' . $fset->name . ' - ' . $fset->display_label . ' ' . $fset->cell_type->name . ' ' . $fset->feature_type->name . ' ' . scalar(@{$fset->get_all_Features}) . ' AnnotatedFeatures
';
    }
}

Example: Get all TFBS FeatureSets

Found 29 CTCF FeatureSets
Nessie_NG_STD_2_ctcf_ren_BR1 - CTCF - IMR90 Enriched Sites
IMR90 CTCF 43427 AnnotatedFeatures
K562_CTCF_ENCODE_Broad_SWEmbl_R015_D150 - CTCF - K562 Enriched Sites
K562 CTCF 32642 AnnotatedFeatures
CD4_CTCF_BarskiZhao_PMID17512414_SWEmbl_R015_D150 - CTCF - CD4 Enriched Sites
CD4 CTCF 25804 AnnotatedFeatures
HepG2_CTCF_ENCODE_Uta_SWEmbl_R015_D150 - CTCF - HepG2 Enriched Sites
HepG2 CTCF 40310 AnnotatedFeatures
GM12878_CTCF_ENCODE_Broad_SWEmbl_R015_D150 - CTCF - GM12878 Enriched Sites
GM12878 CTCF 24837 AnnotatedFeatures
GM12878_CTCF_ENCODE_Uta_SWEmbl_R015_D150 - CTCF - GM12878 Enriched Sites
GM12878 CTCF 33941 AnnotatedFeatures
NHEK_CTCF_ENCODE_Broad_SWEmbl_R015_D150 - CTCF - NHEK Enriched Sites
NHEK CTCF 36209 AnnotatedFeatures
HeLa-S3_CTCF_ENCODE_Uta_SWEmbl_R015_D150 - CTCF - HeLa-S3 Enriched Sites
HeLa-S3 CTCF 42904 AnnotatedFeatures
H1ESC_CTCF_ENCODE_Broad_SWEmbl_R015_D150 - CTCF - H1ESC Enriched Sites
H1ESC CTCF 38179 AnnotatedFeatures
H1ESC_CTCF_ENCODE_Uta_SWEmbl_R015_D150 - CTCF - H1ESC Enriched Sites
H1ESC CTCF 38179 AnnotatedFeatures
HUVEC_CTCF_ENCODE_Broad_SWEmbl_R015_D150 - CTCF - HUVEC Enriched Sites
HUVEC CTCF 31243 AnnotatedFeatures
HepG2_CTCF_ENCODE_Uw_SWEmbl_R015_D150 - CTCF - HepG2 Enriched Sites
HepG2 CTCF 34242 AnnotatedFeatures
GM06990_CTCF_ENCODE_Uw_SWEmbl_R015_D150 - CTCF - GM06990 Enriched Sites
GM06990 CTCF 31001 AnnotatedFeatures
K562_CTCF_ENCODE_Uta_SWEmbl_R015_D150 - CTCF - K562 Enriched Sites
K562 CTCF 42393 AnnotatedFeatures
HeLa-S3_CTCF_ENCODE_Uw_SWEmbl_R015_D150 - CTCF - HeLa-S3 Enriched Sites
HeLa-S3 CTCF 33119 AnnotatedFeatures
HUVEC_CTCF_ENCODE_Uw_SWEmbl_R015_D150 - CTCF - HUVEC Enriched Sites
HUVEC CTCF 36709 AnnotatedFeatures
K562_CTCF_ENCODE_Uw_SWEmbl_R015_D150 - CTCF - K562 Enriched Sites
K562 CTCF 28255 AnnotatedFeatures
NHEK_CTCF_ENCODE_Uw_SWEmbl_R015_D150 - CTCF - NHEK Enriched Sites
NHEK CTCF 29090 AnnotatedFeatures

"n";
Exercise: Set questions

1. DataSets
Datasets are a meta container for data, grouping the data obtained by an analysis (FeatureSets) to the underlying raw data (ResultSets).
• Create a script which fetches all available DataSets for Human.
• How many are there?
• Now get the 'RegulatoryFeatures:MultiCell' data set and print the display label of the product feature set and all the supporting sets.
HINT: Use the DataSetAdaptor methods.

2. FeatureSets
Feature Sets hold processed data or features i.e. peak calls or the output of a high level analysis e.g. the Regulatory Build.
• Print the name of the feature sets for the Human 'GM12878' cell type.
• Print the name of the feature sets for the Human 'CTCF' feature type.
• Is the Human FeatureSet 'VISTA enhancer set' associated to any cell type or feature type?
• Trick question: Get the supporting data for the VISTA FeatureSet.
HINT: Most adaptors have a fetch_by_name method
HINT: DataSetAdaptor->fetch_by_product_FeatureSet will fetch the DataSet containing the supporting/raw data for a FeatureSet.