Babraham Bioinformatics

Analysing High Throughput Sequencing Data with SeqMonk

Version 2018-12

Licence

This manual is © 2008-18, Simon Andrews.

This manual is distributed under the creative commons Attribution-Non-Commercial-Share Alike 2.0 licence. This means that you are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution. You must give the original author credit.
- Non-Commercial. You may not use this work for commercial purposes.
- Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a licence identical to this one.

Please note that:

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

Full details of this licence can be found at
http://creativecommons.org/licenses/by-nc-sa/2.0/uk/legalcode

Table of Contents

Introduction 6
Installing and Configuring SeqMonk 7
Installing the software 7
Windows 7
Max OSX 7
Linux 7
Configuring SeqMonk 8
Initial Configuration 8
Linking to R 9
HTTP Proxy Server 9
Starting a project and Importing Data 10
Creating a new Project 10
Importing Genomes 10
Custom Genomes 11
Importing Extra Annotation 11
Importing Data 12
Generic Text Import 13
Import warnings 14
Organising your Data 15
Renaming Data Sets 15
Grouping Samples 15
Automatic grouping 16
Using the SeqMonk Visualisation Tools 17
Basic Layout 17
The Data Panel 17
The Genome Overview 18
The Chromosome View 18
Using the Chromosome View 20
Changing annotation tracks 20
Changing Data Tracks 20
Moving Around 21
Jumping to a known position 21
Finding Features 22
Changing the Raw Data Display 22
Quantitation 25
Defining Probes 25
Running Window Probe Generator 26
Feature Probe Generator 26
Read Position Probe Generator 26
MACS Peak Caller 26
Current Region Generator 26
Quantitating Data 27
Read Count Quantitation 27
Base Pair Quantitation 28
Enrichment Quantitation 28
Difference Quantitation 28
Percentage Coverage Quantitation 28
Running the Quantitation 28
Viewing Quantitated Data 29
Changing the Data Zoom level 29

Changing the Scale type 29
Decluttering the Display 30
Further Display Options 30
Quantitated Data Figures 31
Distributions - The Probe Value Histogram 31
Distributions - The Probe Length Histogram 32
The Bean Plot 32
The Probe Trend Plot 33
The Scatterplot 33
The Line Graph 34
Requantitating Data 36
Generating a new Probe Set 36
Filtering your Data 37
Applying Filters 37
Filter on Value 37
Filter on Value Differences 38
Filter by Position 38
Filter by Probe Length 38
Filter by Feature 38
Combining Probe Lists 39
Statistical Filters 39
Filter Comments 39
Filter Annotations 39
Exporting Data 40
Exporting Graphics 40
Creating Reports 41
Annotated Probe Report 41
Probe Group Report 42
Feature Report 42
DataStore Summary Report 42
Probe List Description Report 43
Advanced Quantitation 44
Quantitation Pipelines 44
RNA-Seq quantitation pipeline 44
Wiggle plot pipeline 45
Bisulphite methylation over feature 45
Quantitation Normalisation 45
Plotting Quantitation Distributions 46
Applying additional normalisation 47
Match distribution quantitation 47
Percentile normalisation quantitation 48
Enrichment normalisation quantitation 50
Relative quantitation 50
Manual correction quantitation 50
Working with larger data sets 51
Data Store Clustering 51
The Data Store Tree 52
The Correlation Matrix 52
The PCA Plot 52
The Tsne Plot 53
View options for larger studies 54
Different ways to represent Replicate Sets 54

Quantitated data plotting for large sample numbers 55
Chromosome view adjustments for large sample numbers 56
Statistical Filters 57
Running statistical filters 58
Drawing volcano plots for statistical results 58
Filtering on probe list annotation 59
Clustering and Grouping Probes 60
The line graph plot 60
Hierarchical clustering 62
Automated Correlation Clustering 62
Manual clustering 63
Linking SeqMonk to external programs 64
Exporting and importing probes and quantitations 64
Importing hits from an external statistical test 64
Relating probes to an external position list 64
Relating probes to an external feature list 64
Transferring locations between programs 65
Putting quantitated data into another genome browser 65
Quickly matching positions between SeqMonk and a Genome Browser 65

Introduction

High throughput sequencing machines have changed the way we have to think about the analysis of sequencing data. They have the ability to generate tens of millions of sequence reads in a single run, containing hundreds of millions of bases of sequence. Traditional sequence analysis packages were completely unsuited to handling this kind of data and we must therefore use programs which are designed with high-throughput sequencing in mind.

Because high throughput sequencing is a very generic technology it can be applied to many different kinds of experiments and as such there isn't going to be a generic analysis application which will cover all types of experiment. In this course we will concentrate on a common class of experiments where sequence reads are mapped against an annotated genome and the distribution of the mapped fragments is then analysed. This sort of workflow would normally be associated with ChIP-Seq experiments, but works equally well with expression measurement, detection of genomic indels and many other experiment scenarios. This course will not cover experiments where novel SNPs are being sought, or experiments attempting to generate novel assembled sequence since these will require a very different set of software.

SeqMonk is a program developed at the Babraham Institute which was designed to help with the analysis of mapped sequencing data. It was designed when the first high-throughput sequencing data sets started to appear – but can work with any mapped dataset, whatever its source. It combines an annotated genome viewer with a data viewer allowing you to visualise your data against an annotated genome. It also provides tools for quantitating and filtering your data to allow you to find regions of interest. The program provides an environment in which you can easily explore your data and try out a number of different analyses in real time on normal desktop PC hardware.

Installing and Configuring SeqMonk

SeqMonk is free software. It is released under the GPL v3. As such it can be obtained at no cost from:

http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/

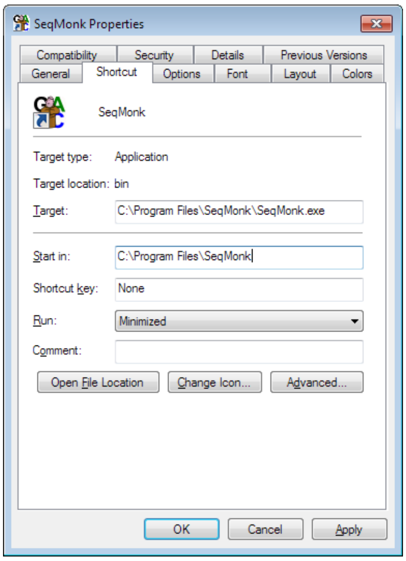
You can also use this address to access all of the documentation for the program as well as an example data set.

If you want to have a look at the SeqMonk source code then this can be found on GitHub at https://github.com/s-andrews/SeqMonk.

Installing the software

A full guide to installing SeqMonk can be found on the downloads page of the project web site, but a summary of the install procedure for all supported platforms is below.

Windows



SeqMonk

Target type: Application

Target location: bin

Target: C:\Program Files\SeqMonk\SeqMonk.exe

Start in: C:\Program Files\SeqMonk

Shortcut key: None

Run: Minimized

Comment:

To install the windows version of SeqMonk simply download the .zip file from the address above and extract its content into a folder. To run the program, you can double click on the seqmonk.exe file. You may find it more convenient to create a shortcut to this file on your desktop. If you do this there are a couple of other options which you can set on the shortcut:

In the Run section you change the option to say "Minimised" then you won't see the black and white command window which runs the program (and which you don't need). You can also press the "Change Icon" button to find a proper SeqMonk icon which you should find in the SeqMonk install folder.

SeqMonk is a java program so you will need to have a java runtime environment installed on your machine. If you don't have one of these you can download one from http://www.java.com. You should be able to use any version of java from 1.6 onwards to run SeqMonk but it's a good idea to always use the latest released version. If you're using a 64-bit version of windows make sure you get the 64-bit version of java too. This may not be the one you're sent to by default and you may need to look at the "all downloads" section of the java.com website.

Max OSX

The Mac version of the program comes as a compressed DMG image. This should be expanded for you automatically when you download it. To install SeqMonk all you need to do is drag the application into your Applications folder or another folder to which you have write access. The OSX version of SeqMonk also relies on an installation of java to work. Specifically, you need to install the 64-bit java development kit (JDK), not just the standard runtime environment (JRE).

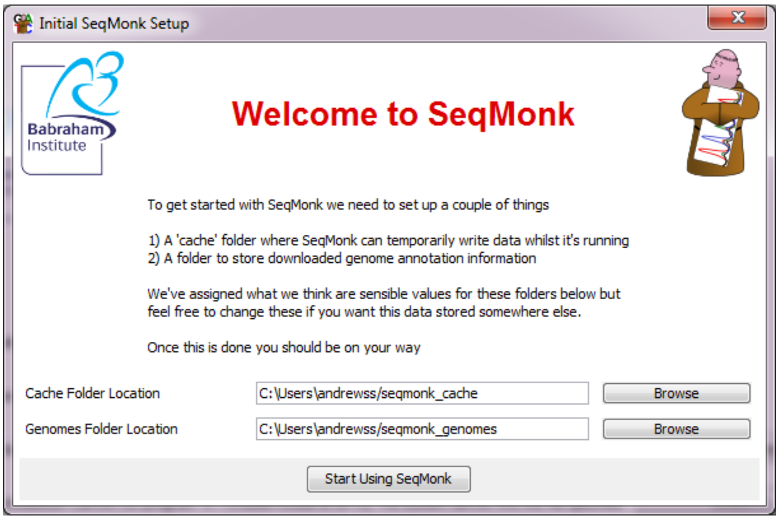
Linux

To install on Linux, you can use the same zip file distribution as for Windows. Expand this into a suitable directory. You can then use the 'SeqMonk' wrapper script to launch the program. Again you will need java to be installed on your system for the program to work. If you want to be able to launch SeqMonk without specifying a full path to the launcher then you can either include the SeqMonk folder in your PATH, or you can symlink (NOT move or copy) the launcher into a directory such as /usr/local/bin/ which is already in your PATH.

Configuring SeqMonk

Initial Configuration

The very first time you run SeqMonk on a machine there are a couple of things you need to set up. To try to make this process as painless as possible the program will guide you through this initial setup. When you first run the program you will see the screen below appear:



To get started with SeqMonk we need to set up a couple of things

1) A 'cache' folder where SeqMonk can temporarily write data whilst it's running
2) A folder to store downloaded genome annotation information

We've assigned what we think are sensible values for these folders below but feel free to change these if you want this data stored somewhere else.

Once this is done you should be on your way

Cache Folder Location: C:\Users\andrewss/seqmonk_cache

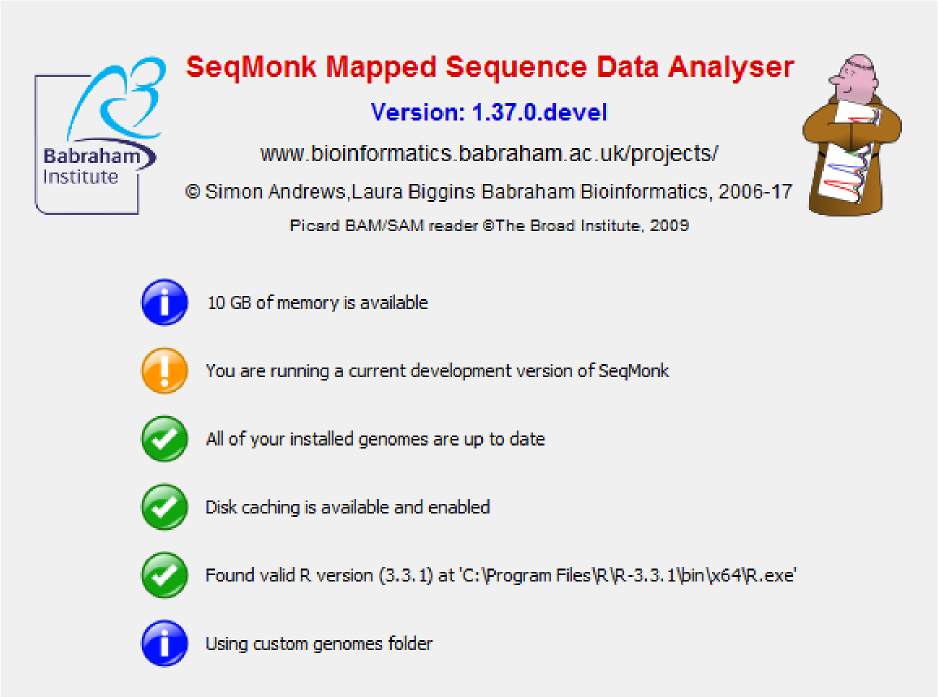
Genomes Folder Location: C:\Users\andrewss/seqmonk_genomes

This screen will configure two folder settings for you. One is a cache folder where SeqMonk will write temporary data whilst it is running. Nothing is permanently stored in this directory so it doesn't need to be backed up, but the volume of data written and read from here can be quite large so you want this folder to be a on a local disk (as opposed to a network drive).

The second folder is a genomes folder which is where SeqMonk will cache copies of the genome annotations for the species and assemblies you are working with. This data will persist on your machine but is just a copy of data on the SeqMonk servers so it doesn't matter if it gets lost.

By default the program will create directories in your home directory for these two locations, but if you want to put them somewhere else you can use the "Browse" buttons to select an alternative location. If you later want to change these settings you can go to Edit > Preferences to change them.

One this initial setup is complete you will see the standard welcome screen which is shown every time SeqMonk is started. This will show some basic information about the install including which version you're running (and if there is an update) – whether there are updates to your installed genomes, and whether you have correctly configured a link to R.


SeqMonk Mapped Sequence Data Analyser

Version: 1.37.0.devel

www.bioinformatics.babraham.ac.uk/projects/

© Simon Andrews,Laura Biggins Babraham Bioinformatics, 2006-17

Picard BAM/SAM reader ©The Broad Institute, 2009

10 GB of memory is available

You are running a current development version of SeqMonk

All of your installed genomes are up to date

Disk caching is available and enabled

Found valid R version (3.3.1) at 'C:\Program Files\R\R-3.3.1\bin\x64\R.exe'

Using custom genomes folder

Linking to R

For some of the statistical and analytical functionality SeqMonk uses a link to a locally installed copy of R to help with the analysis. Creating a link to an R installation is optional and most of SeqMonk will work without it, but you won’t be able to use the filters or views which require an R installation to work.

When you first start SeqMonk it will look to see if it can automatically detect an R installation to use, and whether that installation has all of the packages installed with SeqMonk requires. If no R install is immediately found you can press the “Detect R” button to find it. The program may find it automatically and will offer to use what it found, or you can manually select the folder which contains your R executable.

Once you R installation is set up you many need to install additional R packages which SeqMonk requires. Again SeqMonk will test this automatically when it starts (it takes a few seconds to complete), but if packages are missing then you will have a button which says “Install R dependencies” which you can push to install everything SeqMonk requires.

HTTP Proxy Server

SeqMonk needs to be able to access the internet in order to download genome annotation data and check for updates. If you have a direct connection to the internet, then this will just work. Some sites though force all internet traffic to pass through a proxy server. If you have one of these on your network, you will need to let SeqMonk know where it is. To do this run the SeqMonk program and go to Edit > Preferences and click on the "Network" tab. You can then enter the address of your proxy server and the port it uses. If you don't have a proxy server then just leave the address field blank.

Starting a project and Importing Data

Once you have SeqMonk installed you can create a project and start importing your data.

Creating a new Project

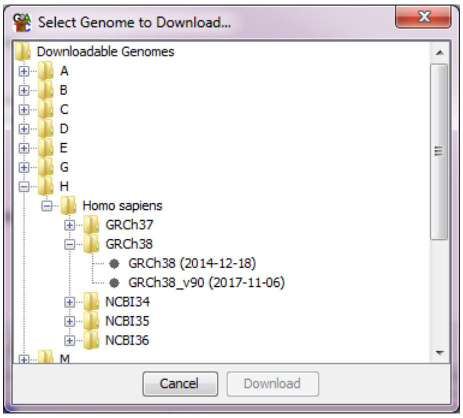
Before you can bring in any data you need to create a new project. The basis for a SeqMonk project is an annotated genome assembly. The genome assembly you use MUST be the same version that your sequencing data was mapped to. If you use the wrong assembly, then any analysis you do will not be valid. If you're not sure which assembly your data was mapped to then you should go back and check before going any further.

To create a new project, select File > New Project. You will then see a list of all of the genomes you currently have installed on your system. If this is the first project you've created, then this list will be empty.

If the genome you want to work with is listed, then you can select it and press OK. If it isn't listed, then you will need to import a genome into the program.

Importing Genomes

SeqMonk can access a large selection of pre-built genome annotations. To see the current list of available genomes, press the "Import Genome From Server" button on the New Project dialog, SeqMonk will then connect to the internet and pull down the list of available genomes.

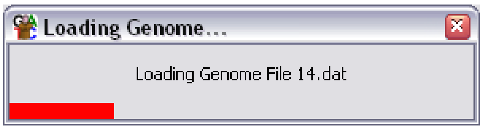


As you can see, the list of genomes is divided alphabetically, and then split by species. Within each species are one or more genome assemblies and for some assemblies there will be different annotation versions. You can select the assembly you want to use and press "Download" this will import the genome onto your machine, and it will then appear on your list of installed genomes when you go to create a new project.

If you don't see the list of available genomes this is normally because SeqMonk had a problem connecting to the internet. Check your internet connection is working, and if it is check that your proxy settings are correct (see the configuration section in the previous chapter).

All core SeqMonk genomes use data taken from the Ensembl project (www.ensembl.org). If you need to know which version of the Ensembl annotation was used for a given assembly, then the date shown in the import dialog will tell you the date on which the genome was processed for use with SeqMonk. More recently processed assemblies will have the Ensembl release version appended to the end of their names (e.g. GRCh38_v90 is annotations from Ensembl release 90 of the GRCh38 human assembly).

Once you have selected the assembly you want to use for your project SeqMonk will open the genome.



Loading Genome File 14.dat

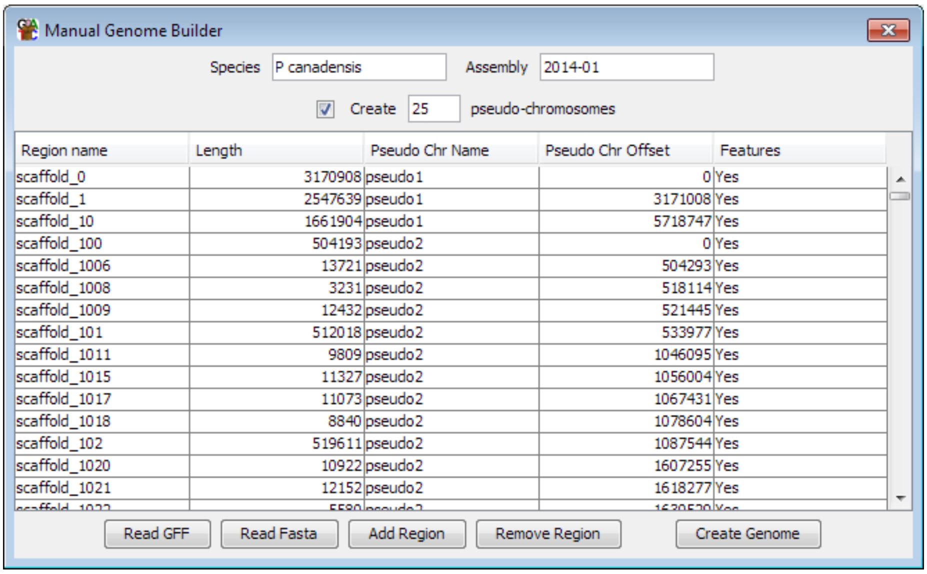
It is really important that you get both the species and the assembly correct. Using the wrong assembly could mean that all your data looks correct, but that reads are mapped incorrectly against the genomic annotations.

When the genome has finished loading SeqMonk will display it and you can begin importing your data.

Custom Genomes

If you are using a genome which isn't listed in the set available from the SeqMonk server then you may need to create a custom genome. If your genome is a chromosome based assembly supported by Ensembl but isn't listed then please simply email babraham.bioinformatics@babraham.ac.uk and we can quickly add it for you. For other genomes you can create a custom genome.

To create a custom genome, you will need either one or more fasta files of sequence or a GTF/GFF file of annotation, and preferably both. To make your custom genome just press the "Build Custom Genome" button in the import dialog and then load in the files you have using the buttons at the bottom of the dialog. You should see the details of your chromosomes appear. You can edit the details of your genome in the table and can choose to selectively delete any sequences you don't want to include.



Species P canadensis Assembly 2014-01

☑ Create 25 pseudo-chromosomes

Region name	Length	Pseudo Chr Name	Pseudo Chr Offset	Features
scaffold_0	3170908	pseudo1	0	Yes
scaffold_1	2547639	pseudo1	3171008	Yes
scaffold_10	1661904	pseudo1	5718747	Yes
scaffold_100	504193	pseudo2	0	Yes
scaffold_1006	13721	pseudo2	504293	Yes
scaffold_1008	3231	pseudo2	518114	Yes
scaffold_1009	12432	pseudo2	521445	Yes
scaffold_101	512018	pseudo2	533977	Yes
scaffold_1011	9809	pseudo2	1046095	Yes
scaffold_1015	11327	pseudo2	1056004	Yes
scaffold_1017	11073	pseudo2	1067431	Yes
scaffold_1018	8840	pseudo2	1078604	Yes
scaffold_102	519611	pseudo2	1087544	Yes
scaffold_1020	10922	pseudo2	1607255	Yes
scaffold_1021	12152	pseudo2	1618277	Yes

If your assembly is scaffold or contig based, then you should probably group your individual sequences into pseudo-chromosomes by ticking the box at the top. This will make the displays in SeqMonk much more manageable and won't affect your analysis.

Once you have created your custom genome you can go back to the original genome selection screen and your genome should be listed.

When using a custom genome, the only limitation is that if you want to transfer a project from one computer to another you will also need to manually copy the custom genome from the genomes folder of the original computer to the equivalent folder on your new computer since SeqMonk won't be able to download this from our servers.

Importing Extra Annotation

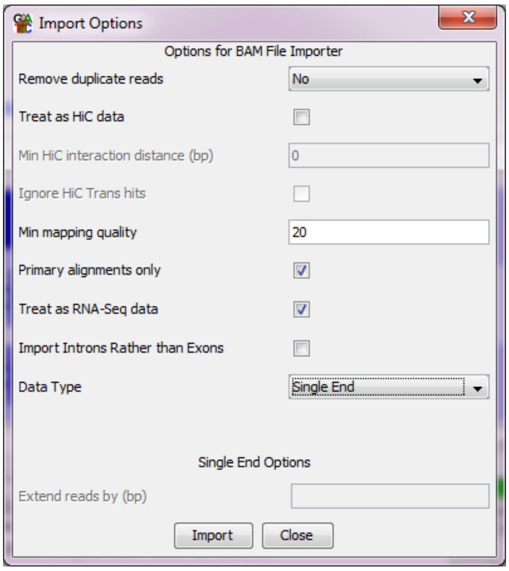
If you have additional annotation information you want to view you can load additional features from text files.

If your file is in GFF or GTF format, then simply select File > Import Annotation > GFF/GTF and choose the file which contains your annotation. You will see a new entry appear in your Annotation Sets folder and you can right-click on this to rename or delete these features. You will be offered the option of prefixing the feature names with some text you specify so that you could, for example, tell genes from your GFF file apart from genes which came from the core annotation.

If your annotations are in a non-standard format, then you can use the generic text import feature to bring them in. Select File > Import Annotation > Text (Generic). You can select your file and then use the dialog which opens to say which delimiter splits the fields in your file and which columns contain which pieces of information. The only columns you need to import are start, end and chromosome. All of the others are optional.

You new features will immediately become available in your genome view and can be added and removed in the same way as any other feature track.

Importing Data


Import Options

Options for BAM File Importer

Remove duplicate reads: No

Treat as HiC data: □

Min HiC interaction distance (bp): 0

Ignore HiC Trans hits: □

Min mapping quality: 20

Primary alignments only: ☑

Treat as RNA-Seq data: ☑

Import Introns Rather than Exons: □

Data Type: Single End

Single End Options

Extend reads by (bp):

The only information SeqMonk needs to import your data is a series of mapped genomic positions. These comprise a chromosome, start and end positions and optionally a strand for each read. Most mapped data will be distributed in BAM format, but SeqMonk understands many additional formats and even has a generic text import option so you can bring in data from any delimited text file.

To import your data simply select File > Import Data and select the format of file you want to import. For any standard format there are a couple of options you may be offered. The program will examine your data and will try to set the correct options automatically, but you can change these if you prefer.

- Single / Paired End import. If you have generated paired end sequencing data, then you can choose to import this either as single ended or paired end data. If you import as single end data then two reads will show up for every sequence, one at each end, but there will be no connection between them. If you choose paired end data, then SeqMonk will join together the two ends of each sequence into a single extended read. The strand of this paired read will be the strand of the earlier mapping sequence. If you paired reads are too far apart, or are on different chromosomes then they are assumed to have been mis-mapped and are discarded.

- Remove Duplicates. You can choose to only import unique reads from each data set. Any duplicates will be removed at the import stage and won't show up in the program. You will have the option to ignore duplicates later when quantitating your data, so you generally don't want to remove them at the import stage. If you do remove duplicates you can choose whether a duplicate is defined by its start position, end position or both start and end (usually just using the start position is the right thing to do for single end data, and using both ends is correct for paired end data).

- Import primary alignments. If your mapper has allowed a single sequence to be mapped to more than one location, then one of the mapped positions will be flagged as a primary alignment. By default, SeqMonk will only import primary alignments so that a maximum of one position is imported for each read in your original data file. If you want to import all positions for multiply mapped reads, then you can untick this box to get all of them.

- Filter on mapping quality. When a read mapper generates an alignment to the genome it gives it a mapping quality (MAPQ) score. The meaning of this score is different for each aligner, but in general a lower score reflects a higher number of mismatches between the read and the reference, and a lack of uniqueness in the hit. You will normally therefore want to filter your incoming reads to keep only the unique high identity alignments, and setting a threshold for the quality scores allows you to do this. SeqMonk defaults to a

MAPQ cut-off of 20 which is sensible for most aligners but you should look at the documentation for the aligner you used to be sure.

- Extending single end reads. If you are importing a single end ChIP-seq dataset, then you may want to use the extend reads option. This extends each read by a fixed size so that your read is more realistically positioned over the fragment you enriched, rather than just showing a short tag from the end. This makes the downstream analysis easier since you can see more realistic peaks over areas of enrichment. The amount you extend by should reflect the average size of the fragments which went into your library (minus the read length of your sequences).

- RNA-Seq import. If you have done a spliced mapping you will be offered the option to treat the file as RNA-Seq data and split spliced reads into their component parts. This will generate multiple reads for each mapped read to cover the different spliced portions. If you do not select this option, then the read will cover the whole mapped region – but may be rejected if the mapped region is too long (which can happen if the spliced out region is long). If you choose to do this, you can then further opt to import the introns instead of the exons. This would be useful if you wanted to analyse splicing in RNA-Seq data. If you opt to import RNA-Seq reads then you can also select an additional option to import introns instead of exons if you want to analyse splicing.

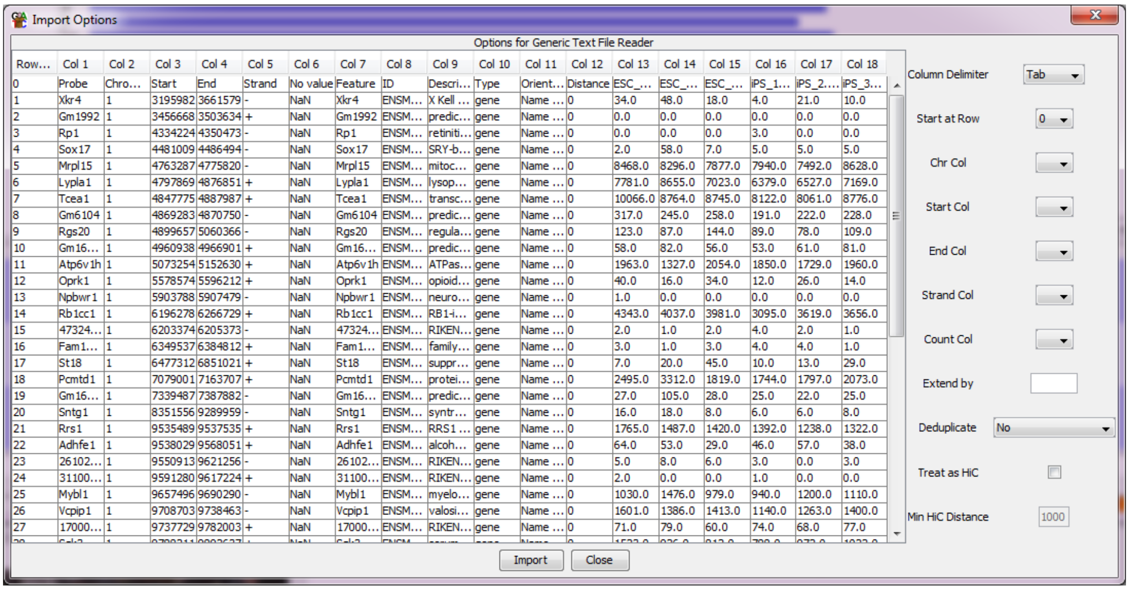
- HiC import. If you have a dataset consisting of HiC pairs then you can tell the importer to treat alternating sequences as HiC data. You need to ensure that the whole file contains paired sequences, with no single sequences otherwise the pairing will not be correct. You can also choose to ignore interactions smaller than a given distance.

For BAM/SAM files you should find that the program tries to auto-detect the pairing and splicing settings, so for the most part you shouldn't need to change the import options. You will need to pay attention to the options to extend reads, or to treat your data as HiC since these cannot be determined automatically.

If you choose to do a generic text import, then there are a number of different options you need to set.

Generic Text Import

You can import multiple files in one Generic Text Import but they must all have the same format. Once you have selected the files you want to import you will see a preview of your data and you will need to answer some questions about it:



Row...	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Col 11	Col 12	Col 13	Col 14	Col 15	Col 16	Col 17	Col 18
0	Probe	Chro...	Start	End	Strand	No value	Feature	ID	Descri...	Type	Orient...	Distance	ESC_...	ESC_...	ESC_...	iPS_1...	iPS_2....	iPS_3...
1	Xkr4	1	3195982	3661579	-	NaN	Xkr4	ENSM...	X Kell ...	gene	Name ...	0	34.0	48.0	18.0	4.0	21.0	10.0
2	Gm1992	1	3456668	3503634	+	NaN	Gm1992	ENSM...	predic...	gene	Name ...	0	0.0	0.0	0.0	0.0	0.0	0.0
3	Rp1	1	4334224	4350473	-	NaN	Rp1	ENSM...	retiniti...	gene	Name ...	0	0.0	0.0	0.0	3.0	0.0	0.0
4	Sox17	1	4481009	4486494	-	NaN	Sox17	ENSM...	SRY-b...	gene	Name ...	0	2.0	58.0	7.0	5.0	5.0	5.0
5	Mrpl15	1	4763287	4775820	-	NaN	Mrpl15	ENSM...	mitoc...	gene	Name ...	0	8468.0	8296.0	7877.0	7940.0	7492.0	8628.0
6	Lypla1	1	4797869	4876851	+	NaN	Lypla1	ENSM...	lysop...	gene	Name ...	0	7781.0	8655.0	7023.0	6379.0	6527.0	7169.0
7	Tcea1	1	4847775	4887987	+	NaN	Tcea1	ENSM...	transc...	gene	Name ...	0	10066.0	8764.0	8745.0	8122.0	8061.0	8776.0
8	Gm6104	1	4869283	4870750	-	NaN	Gm6104	ENSM...	predic...	gene	Name ...	0	317.0	245.0	258.0	191.0	222.0	228.0
9	Rgs20	1	4899657	5060366	-	NaN	Rgs20	ENSM...	regula...	gene	Name ...	0	123.0	87.0	144.0	89.0	78.0	109.0
10	Gm16...	1	4960938	4966901	+	NaN	Gm16...	ENSM...	predic...	gene	Name ...	0	58.0	82.0	56.0	53.0	61.0	81.0
11	Atp6v1h	1	5073254	5152630	+	NaN	Atp6v1h	ENSM...	ATPas...	gene	Name ...	0	1963.0	1327.0	2054.0	1850.0	1729.0	1960.0
12	Oprk1	1	5578574	5596212	+	NaN	Oprk1	ENSM...	opioid...	gene	Name ...	0	40.0	16.0	34.0	12.0	26.0	14.0
13	Npbwr1	1	5903788	5907479	-	NaN	Npbwr1	ENSM...	neuro...	gene	Name ...	0	1.0	0.0	0.0	0.0	0.0	0.0
14	Rb1cc1	1	6196278	6266729	+	NaN	Rb1cc1	ENSM...	RB1-i...	gene	Name ...	0	4343.0	4037.0	3981.0	3095.0	3619.0	3656.0
15	47324...	1	6203374	6205373	-	NaN	47324...	ENSM...	RIKEN...	gene	Name ...	0	2.0	1.0	2.0	4.0	2.0	1.0
16	Fam1...	1	6349537	6384812	+	NaN	Fam1...	ENSM...	family...	gene	Name ...	0	3.0	1.0	3.0	4.0	4.0	1.0
17	St18	1	6477312	6851021	+	NaN	St18	ENSM...	suppr...	gene	Name ...	0	7.0	20.0	45.0	10.0	13.0	29.0
18	Pcmtd1	1	7079001	7163707	+	NaN	Pcmtd1	ENSM...	protei...	gene	Name ...	0	2495.0	3312.0	1819.0	1744.0	1797.0	2073.0
19	Gm16...	1	7339487	7387882	-	NaN	Gm16...	ENSM...	predic...	gene	Name ...	0	27.0	105.0	28.0	25.0	22.0	25.0
20	Sntg1	1	8351556	9289959	-	NaN	Sntg1	ENSM...	syntr...	gene	Name ...	0	16.0	18.0	8.0	6.0	6.0	8.0
21	Rrs1	1	9535489	9537535	+	NaN	Rrs1	ENSM...	RRS1 ...	gene	Name ...	0	1765.0	1487.0	1420.0	1392.0	1238.0	1322.0
22	Adhfe1	1	9538029	9568051	+	NaN	Adhfe1	ENSM...	alcoh...	gene	Name ...	0	64.0	53.0	29.0	46.0	57.0	38.0
23	26102...	1	9550913	9621256	-	NaN	26102...	ENSM...	RIKEN...	gene	Name ...	0	5.0	8.0	6.0	3.0	0.0	3.0
24	31100...	1	9591280	9617224	+	NaN	31100...	ENSM...	RIKEN...	gene	Name ...	0	2.0	0.0	0.0	1.0	0.0	0.0
25	Mybl1	1	9657496	9690290	-	NaN	Mybl1	ENSM...	myelo...	gene	Name ...	0	1030.0	1476.0	979.0	940.0	1200.0	1110.0
26	Vcpip1	1	9708703	9738463	-	NaN	Vcpip1	ENSM...	valosi...	gene	Name ...	0	1601.0	1386.0	1413.0	1140.0	1263.0	1400.0
27	17000...	1	9737729	9782003	+	NaN	17000...	ENSM...	RIKEN...	gene	Name ...	0	71.0	79.0	60.0	74.0	68.0	77.0

Column Delimiter: Tab

Start at Row: 0

Chr Col:

Start Col:

End Col:

Strand Col:

Count Col:

Extend by:

Deduplicate: No

Treat as HiC ☐

Min HiC Distance: 1000

The main table is a preview of your data, split into columns. On the right are the boxes you need to fill in. What you need to tell the program is:

- What delimiter is present between the different fields on each line of your file
- On which row your actual data starts (so you can remove any headers)
- Which columns in your file contain the data for
  - The chromosome
  - The start position
  - The end position
  - The strand (this column is optional)
  - The count (this column is optional)

Column Numbers are shown above your data, and the first column you see contains the row numbers.

Once you have specified all of this information you should be able to press the "Import" button to import your data.

Import warnings

After you have finished importing your data you may see a list of warnings which were generated whilst importing your data. You may consider some of these warnings to be harmless, but you should check through them since they could potentially indicate a serious problem with your data.

Some of the more commonly seen warnings are listed below:

- Couldn't find a chromosome called 'X'. If a read uses a chromosome name which couldn't be understood, then you will see this message. In some cases, this will be because the nomenclature differs between your mapping program and SeqMonk (e.g. 2 vs II or M vs Mt), in other cases the chromosome name may be part of a longer string which SeqMonk can't interpret (it tries lots of different ways to figure these out), or may be just an accession code rather than a chromosome name. If you don't care about the data you've lost (maybe just the mitochondrion) then you can ignore this. Alternatively, you can edit a file called aliases.txt in the assembly folder of your genomes folder to add in name mappings which the program can't figure out for itself.

- Read position was beyond the end of the chromosome. If you have a chromosome which is 10kb long, then a read whose position was 11kb will be rejected. If you see these messages, then this is often an indication that you have selected the wrong assembly. If this is the case, then you need to stop and start your project again using the correct assembly. In some cases, this message can be harmless though – if you're using an assembly which has a circular genome then you may find that reads are mapped slightly off each end of a chromosome, but only by a few bases. Also if you've chosen to extend your reads then they may fall a few bases over the end of the chromosome. Any more than that and this message indicates a potentially serious problem

- For paired end data you will probably see warnings for read pairs which were mapped further apart than the limit which you set during the import setup. This can occur in a small number of cases due to mis-mapping of one end, and is harmless as long as the proportion of the library affected is low.

Organising your Data

Renaming Data Sets

When you initially import your data your samples will be named using the filename from which the data came. If you want to change this to something more relevant, then you can do this by selecting Data > Edit DataSets. If you select the sample, you want to rename you can then press "Rename DataSet" and put in the name you want to use. Alternatively, you can right-click on the data view and select 'Rename'.

Where you need to apply a similar change to a large number of data sets you can go to Data > Edit Data Sets where there is a find/replace option which can do simple (text string replacement) or complex (regular expression replacement) changes to large sets of data sets at once.

If you need to do a bulk renaming of your datasets then you can select multiple sets and use the “Rename DataSet” button to bring up a small table into which you can paste a set of new names to do all of the renaming in a single step.

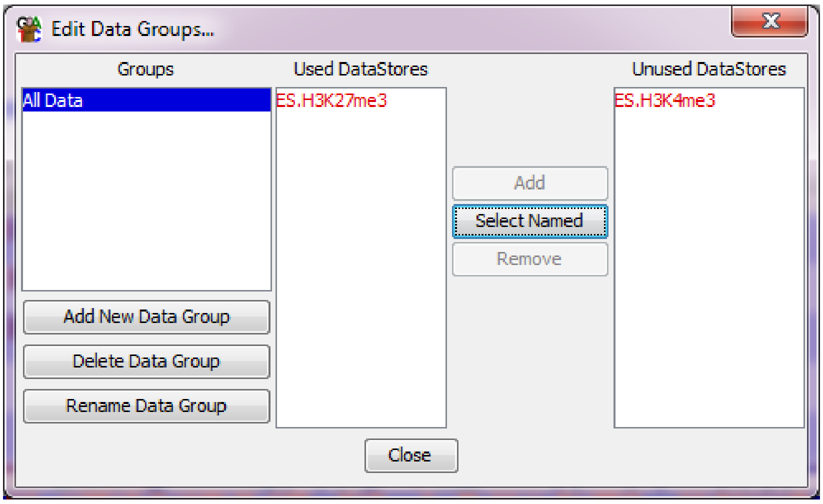
Grouping Samples

In some cases, it will make sense for you to group together the raw data you have imported into larger groups. SeqMonk supports two different kinds of grouping which are appropriate for different kinds of experiments.

- **Data Groups** are a way to combine raw data sets so that they appear as one big data set. Data groups hold their own quantitation values and are intended to be used where you have multiple raw data files coming from the same library.

- **Replicate Sets** are a way to group together data sets or data groups which come from different replicates of the same biological condition. They do not hold their own quantitated values, but instead display the mean value of the data stores they contain. They are the basis for some of the statistical analyses which rely on knowing about replicates.

The process for creating data groups and replicate sets is essentially the same. To start you select Data > Edit Groups, or Data > Edit Replicate Sets from the main menu:



You can create new groups using the button on the left. Once you have created a group you can add samples to it by selecting it and then picking one or more datasets from the 'Unused DataSets' list and pressing 'Add'. You can put the same data set into more than one group if this makes sense for your experiment.

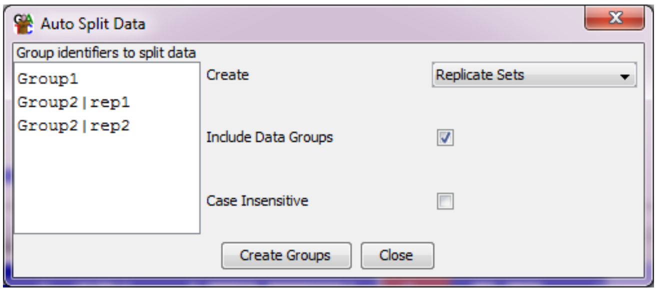
Using this same screen, you can delete groups you no longer need (this will NOT delete the underlying data – just the group), and you can rename existing groups.

It's a good idea to create your groups before you start to do any quantitation or analysis of your data.

Automatic grouping

If you have large numbers of data sets which need to be assembled into Data Groups or Replicate Sets and they use a consistent naming scheme, you can use this to automate the grouping. If your sample names are informative to say which group the sample should belong to then, instead of having to manually create replicate sets you can do this in an automated way based on the names of your samples.

The option to do this is Data > Auto Create Groups/Sets.



Auto Split Data

Group identifiers to split data

```
Group1
Group2|rep1
Group2|rep2
```

Create: Replicate Sets

Include Data Groups ☑

Case Insensitive □

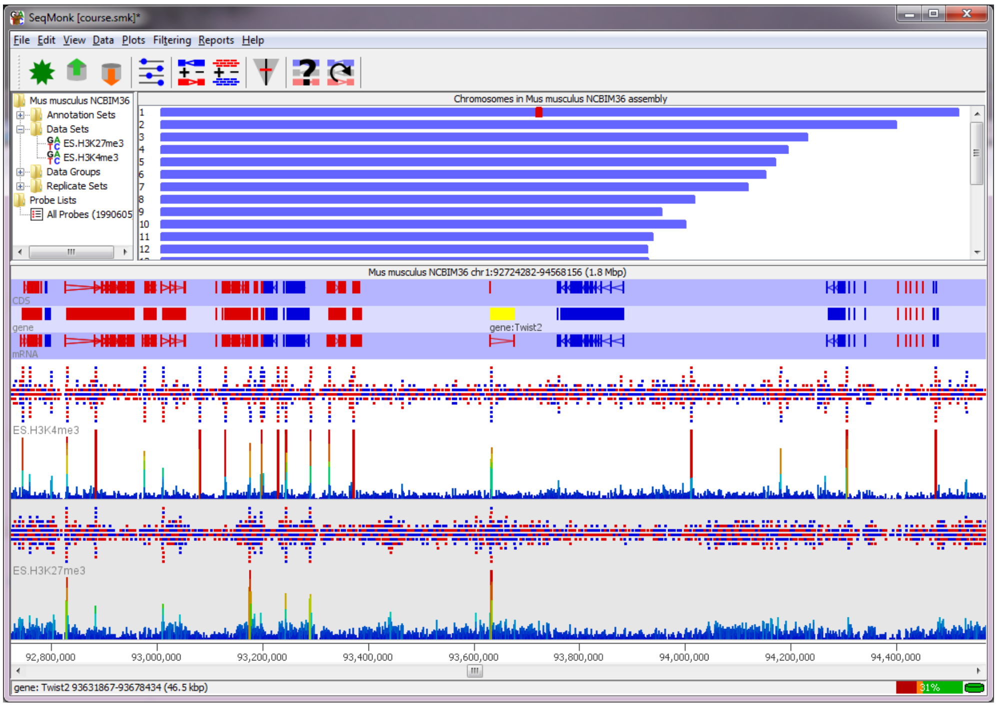
In the box on the left you can put in a set of text strings which can be found in the names of your data sets and which mark the groups you want to create. The tool will then look for a match to each string and will make either a data group or a replicate set out of the samples which match. If you need to use more than one discontinuous string match to define your groups, then you can separate strings with bar characters to make more complex patterns. Groups made this way can always be edited with the traditional data group and replicate set editors.

When creating groups this way, each data set will only be placed into one group, and if multiple matches exist for that dataset then only the first one will be used.

Using the SeqMonk Visualisation Tools

A lot of the time you spend using SeqMonk will be spent looking at your data in different parts of the genome using the built in visualisation tools. Before you worry about quantitating or analysing your data it's therefore worth getting used to the visualisation interface which SeqMonk provides.

Basic Layout



The SeqMonk interface is laid out in panels separated by dividers which you can move to suit your taste. The main panels are:

- The Data Panel (top left)
- The Genome Overview (top right)
- The Chromosome View (bottom)

The Data Panel

The data panel contains a series of folders in which you can see the different annotation and data sets you have imported and the data and replicate groups you have created. Once you've started doing some quantitation then you will also start to see probe lists in this panel.

You can use the data panel to change the properties of any data sets / group. If you select a dataset or group in the data panel it will become highlighted in the chromosome view if it is visible. If you right click on a dataset or group, you will get a small menu which allows you to rename that sample or to turn it off or on in the chromosome view. There are also a couple of plots which you can draw.

Once you have done some quantitation you can select probe subsets in the data panel. This will filter the chromosome view to show you only the probes which are present in that subset.

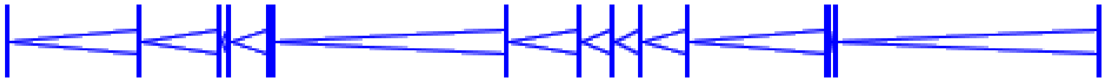
The Genome Overview

The genome overview shows you a graphical representation of your whole genome laid out in chromosomes. On one of the chromosome you should see a red box. This box indicates the region which is currently visible in the chromosome view. When you've quantitated your data and have selected a single sample in the data view you will see a summary of the quantitation across the genome, but until then you'll just see plain blue boxes.

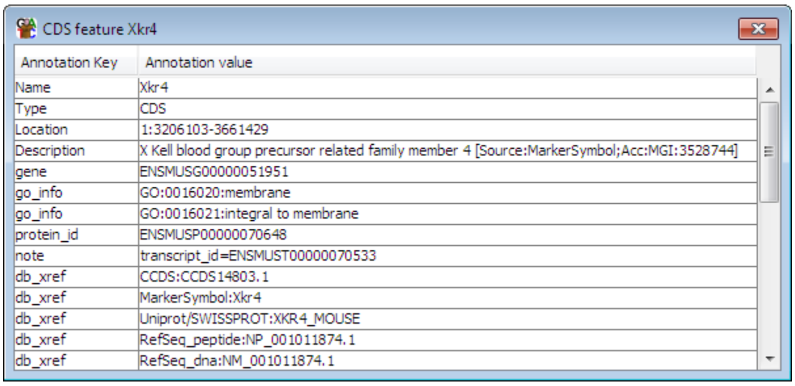
If you click and drag anywhere on the genome view you will move the chromosome view to show the region you selected.

The Chromosome View

The chromosome view is the most detailed view of the genome, and is also where you will focus your attention most of the time you're using SeqMonk. The chromosome view is arranged into tracks. At the top of the view are a series of tracks with blue backgrounds (alternating light and dark) which contain annotation information for different classes of feature (genes, CDS, mRNA etc.). Features appear as blocks which are coloured according to their orientation (red for forward and blue for reverse). Where a feature has multiple linked locations (e.g. exons in a transcript) the blocks are joined by line arrows which also indicate the direction of the feature.



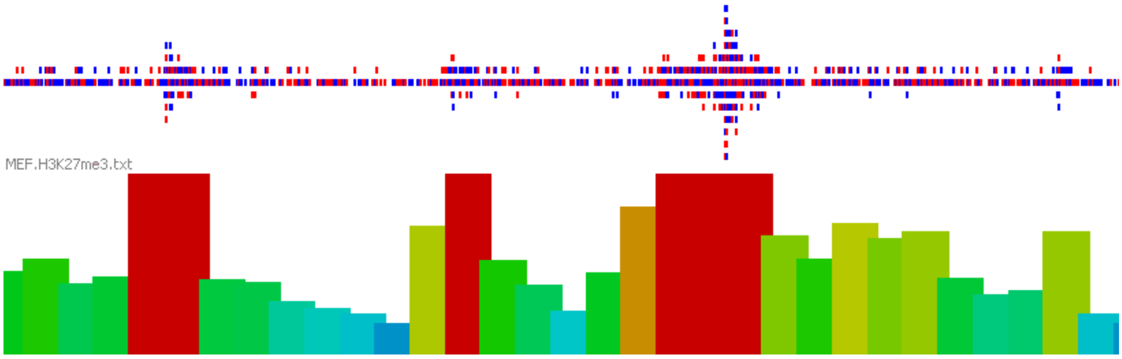
If you put your mouse over a feature you should see a see a label appear under it which shows its name and the type of the feature (e.g.: CDS:Sntg1). If your tracks are quite compressed there may not be enough room for these labels, but this information will still appear in the status bar at the bottom of the program. If you double click on a feature you will get a new window appear containing all of the annotation information for that feature.



CDS feature Xkr4

Annotation Key	Annotation value
Name	Xkr4
Type	CDS
Location	1:3206103-3661429
Description	X Kell blood group precursor related family member 4 [Source:MarkerSymbol;Acc:MGI:3528744]
gene	ENSMUSG00000051951
go_info	GO:0016020:membrane
go_info	GO:0016021:integral to membrane
protein_id	ENSMUSP00000070648
note	transcript_id=ENSMUST00000070533
db_xref	CCDS:CCDS14803.1
db_xref	MarkerSymbol:Xkr4
db_xref	Uniprot/SWISSPROT:XKR4_MOUSE
db_xref	RefSeq_peptide:NP_001011874.1
db_xref	RefSeq_dna:NM_001011874.1

Underneath the feature tracks are a series of white/grey data tracks and each of these will contain the data for one Data Set or Data Group. The information shown in these tracks can be changed, but they can show either raw read positions or quantitated values calculated from the reads.



Where both the reads and quantitated data are shown the reads will be on the top. Reads are coloured according to the stand they map to (red=forward, blue=reverse, grey=unknown) and are packed together as tightly as possible. Where reads would overlap they are spread out vertically to use all of the available space. If there are too many reads for a position than can be stacked in the display, then the extra reads are omitted.

Quantitated data is displayed as a column graph underneath the reads. The colours of the bars are scaled from cold (blue) to medium (green) to hot (red), so the colours and the heights of the bars show the same information.

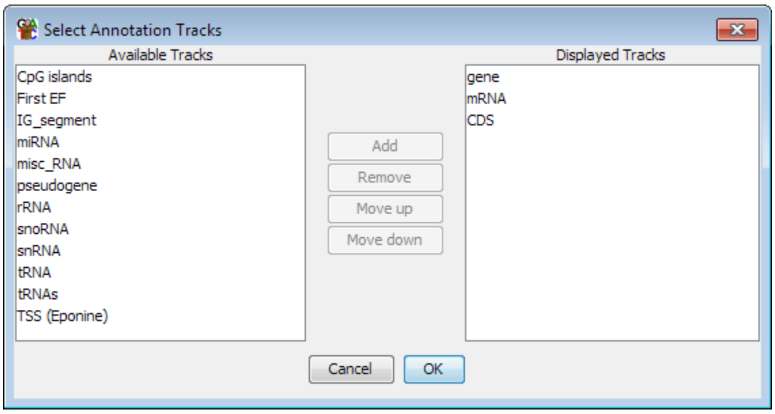
Using the Chromosome View

The chromosome view is a dynamic window in that you can change the information it displays and you can zoom in and out of it to see exactly the region you want.

It is quite common for a SeqMonk project to contain enough different data sets and annotation features that you won't want to view all of them all of the time. You can therefore modify the list of visible tracks to show you only what you need to see.

Changing annotation tracks

To change the genome annotation tracks shown you need to select View > Set Annotation Tracks.



Available Tracks:
CpG islands
First EF
IG_segment
miRNA
misc_RNA
pseudogene
rRNA
snoRNA
snRNA
tRNA
tRNAs
TSS (Eponine)

Displayed Tracks:
gene
mRNA
CDS

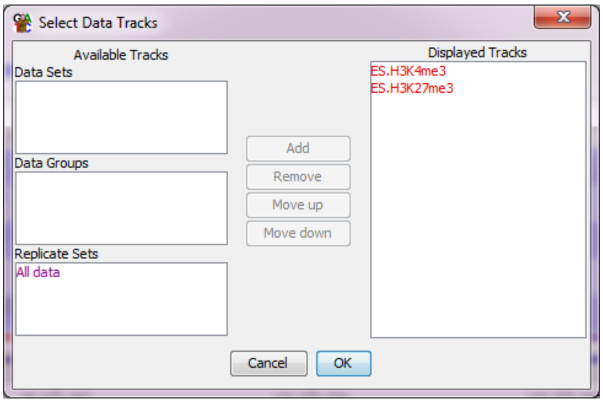
On the left is a list of annotation tracks which have been loaded, but are not currently displayed, on the right is a list of the currently displayed tracks. You can use the buttons in the middle to move tracks from one list to the other and to change the order of tracks within the displayed tracks list.

This dialog will only show tracks which SeqMonk loaded from the genome, but there may be other tracks which are available. Because annotation tracks consume a lot of memory

Changing Data Tracks

You can use a similar approach to change the set of data tracks displayed by SeqMonk. In the main menu you select View > Set Data Tracks.

Again, on the left is a list of data sets, groups and replicate sets which are loaded but not currently displayed. On the right is the list of currently displayed tracks. You use the buttons in the middle to move tracks from one side to the other, or to adjust the order of the displayed tracks.



Replicate Sets:
All data

Displayed Tracks:
ES.H3K4me3
ES.H3K27me3

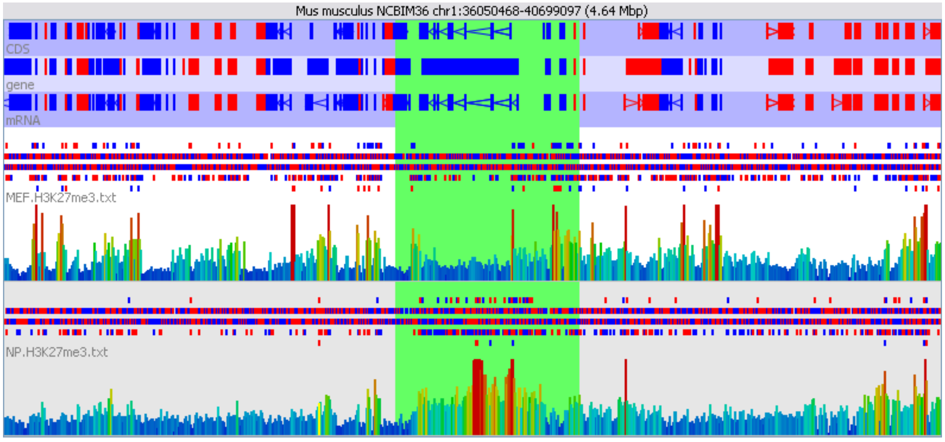
In addition to using this tool you can also add and remove data tracks using the data display. If you right click on any data set or data group, then you will see a menu on which the top entry says "Show track in chromosome view". If this entry has a tick next to it then the track is already visible. Selecting this menu option will toggle the visibility of that track. If you add a track this way, it is always added to the bottom of your current set of tracks.

Moving Around

There are a few different ways to move around the chromosome view.

The quickest way to move along the chromosome view is to use the scroll bar underneath it. If you drag this, you can quickly scan from one end of the chromosome to the other. For more fine-grained movements you might find it easier to use the left and right keys on your keyboard which will move the view in small increments. You can also scroll the scroll wheel on your mouse to make small adjustments to your position within a chromosome.

To zoom in and out of the view it is probably best to use the mouse. To zoom in all you need to do is to click and drag within the chromosome view. You will see that as you drag a green box will shade the region you've selected.



When you release the mouse the display will zoom in to show just the region you selected.

To zoom out you simply press the right mouse button (apple+click on a mac) and the display will stay centred at the same point, but will show twice as much of the genome. You may find that instead of scrolling laterally across the genome it ends up being easier to zoom out to locate the region you're after, and then zoom back in again.

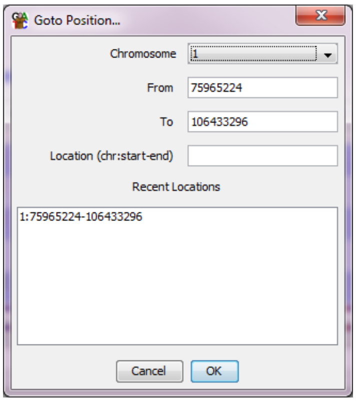
If you want to you can also zoom using the up and down arrows on your keyboard.

Jumping to a known position

If you know exactly where you want to go then rather than having to use the normal movement controls to get there you can jump straight to that position. You can do this by selecting Edit > Goto Position (or press Ctrl+G).

You can then select the chromosome you want and specify a start and end position and the display will then move to that exact position. You can also enter a string of the form chr: start-end in the location box.

The Goto dialog will remember the last few places you visited in the genome so you can easily move back to somewhere you were recently.


Goto Position...

Chromosome 1

From 75965224

To 106433296

Location (chr:start-end)

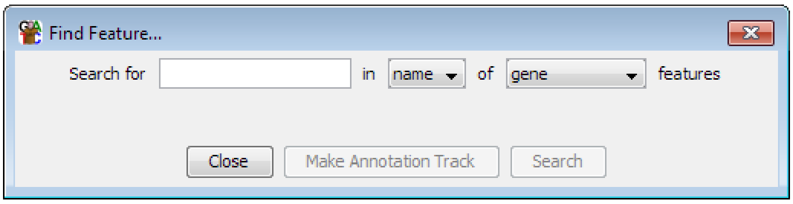
Recent Locations

1:75965224-106433296

Finding Features

If you want to jump straight to a particular feature, then the easiest way to do this is to do a search for the feature and then let SeqMonk take you there directly.

To start a feature search you select Edit > Find Feature (or press Ctrl+F).

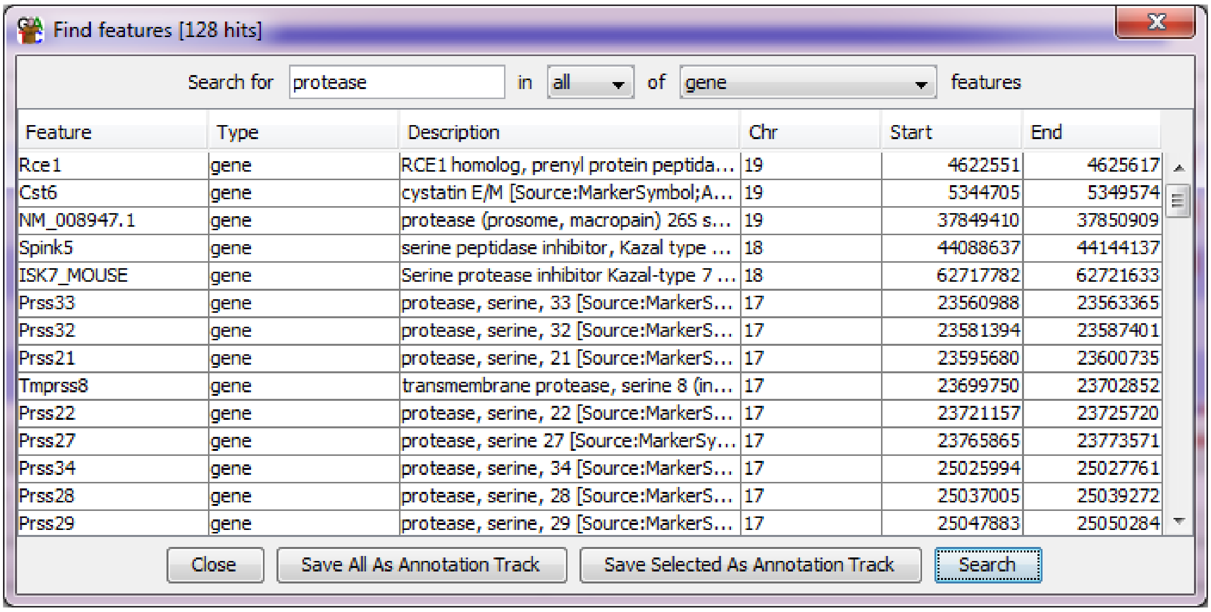


Find Feature...

Search for [ ] in name of gene features

You can choose what type of feature you want to search for and then supply some search terms you're interested in. You can search either just within the name of the feature or through all of the annotation for every feature. By default, SeqMonk only loads a name and a description for each feature. If you want more detail you can enable this by ticking the “Load full genome annotation” box under Edit > Preferences > Memory, although this will take up more memory to store this data. Once the full annotation is loaded it is searchable by the find feature tool.

Once you've done a search you will see the list of hits the program found.



Find features [128 hits]

Search for protease in all of gene features

Feature	Type	Description	Chr	Start	End
Rce1	gene	RCE1 homolog, prenyl protein peptida…	19	4622551	4625617
Cst6	gene	cystatin E/M [Source:MarkerSymbol;A…	19	5344705	5349574
NM_008947.1	gene	protease (prosome, macropain) 26S s…	19	37849410	37850909
Spink5	gene	serine peptidase inhibitor, Kazal type …	18	44088637	44144137
ISK7_MOUSE	gene	Serine protease inhibitor Kazal-type 7 …	18	62717782	62721633
Prss33	gene	protease, serine, 33 [Source:MarkerS…	17	23560988	23563365
Prss32	gene	protease, serine, 32 [Source:MarkerS…	17	23581394	23587401
Prss21	gene	protease, serine, 21 [Source:MarkerS…	17	23595680	23600735
Tmprss8	gene	transmembrane protease, serine 8 (in…	17	23699750	23702852
Prss22	gene	protease, serine, 22 [Source:MarkerS…	17	23721157	23725720
Prss27	gene	protease, serine 27 [Source:MarkerSy…	17	23765865	23773571
Prss34	gene	protease, serine, 34 [Source:MarkerS…	17	25025994	25027761
Prss28	gene	protease, serine, 28 [Source:MarkerS…	17	25037005	25039272
Prss29	gene	protease, serine, 29 [Source:MarkerS…	17	25047883	25050284

Close
Save All As Annotation Track
Save Selected As Annotation Track
Search

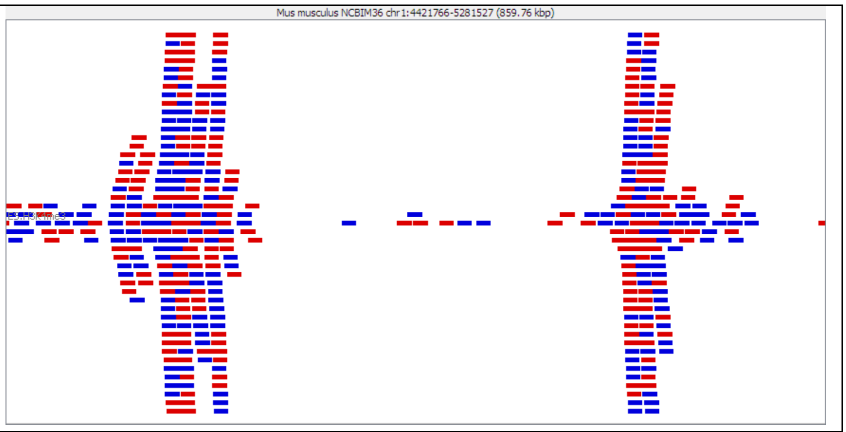
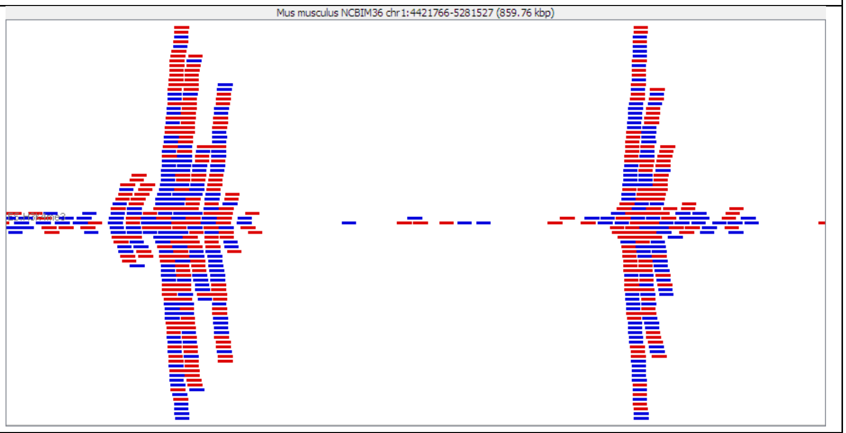
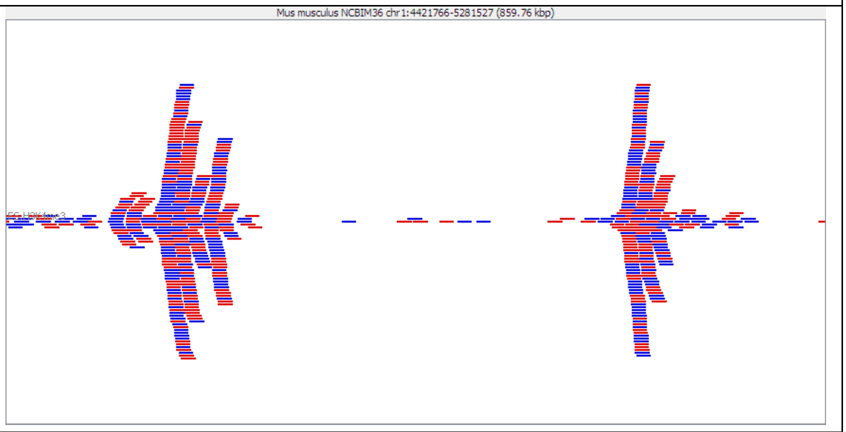
You can sort this list by clicking on any of the column headers. If you want to see a feature in the chromosome view, then you simply need to double click on any line in the results and you will be taken there. If you want to keep hold of this list of features, then you can press the "Make annotation track" button to create a new annotation track containing only these features.

Changing the Raw Data Display

There are many options for changing the way your data is displayed in the chromosome view. All of these are collected together under a single button on the toolbar which shows all of the display preferences options.

When you are viewing the raw sequence data in SeqMonk there are a couple of options which might prove useful. The default display will show a low density of reads and will pack together the forward and reverse reads to make the best use of the available space.

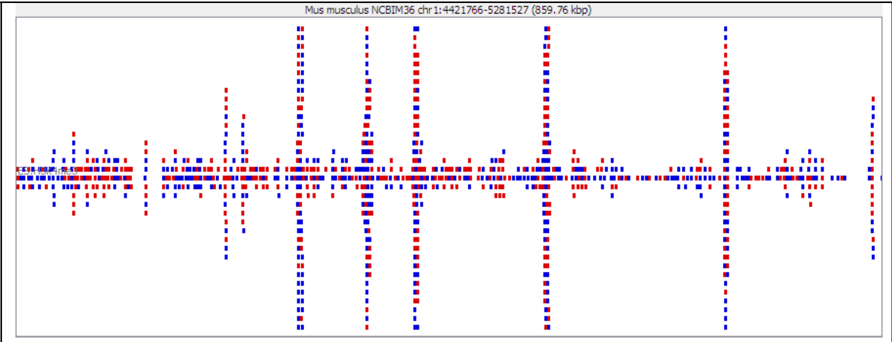
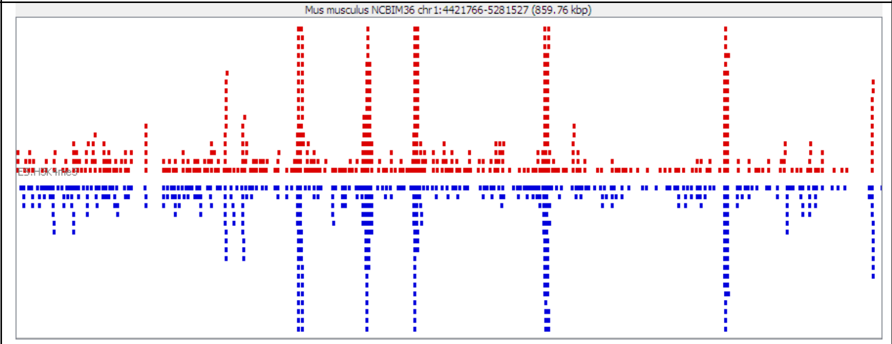
If you have a large amount of data to display, or you have a lot of data sets visible at the same time you might want to increase the density with which the reads are displayed to allow you to see more data at once.

Low Density	 Mus musculus NCBIM36 chr1:4421766-5281527 (859.76 kbp)
Medium Density	 Mus musculus NCBIM36 chr1:4421766-5281527 (859.76 kbp)
High Density	 Mus musculus NCBIM36 chr1:4421766-5281527 (859.76 kbp)

The read density can be set from the display preferences by changing the “Raw Read Display Density” option.

The other display option for raw reads is to separate out the forward and reverse reads from each other and put them in different parts of the track. Depending on the nature of your library the direction of your read may or may not be important. If you pack your reads together then in many cases you are only looking at a summarised representation of your data, and you can't rely on the overall colour of the summary reflecting accurately the read direction composition. In these cases, you can choose to pack forward and reverse reads separately so you can clearly see the breakdown of forward and reverse reads in each region.

You can change the packing through the display preferences by changing the “Display Reads with” option.

Together	 Mus musculus NCBIM36 chr1:4421766-5281527 (859.76 kbp)
Separately	 Mus musculus NCBIM36 chr1:4421766-5281527 (859.76 kbp)

Quantitation

As well as showing raw data SeqMonk is able to quantitate that data. This allows you to compare different conditions or to filter out regions of interest.

Quantitation in SeqMonk traditionally happens in two steps.

1. You define a set of 'probes'. A probe is simply a region of the genome in which you intend calculating a quantitated value

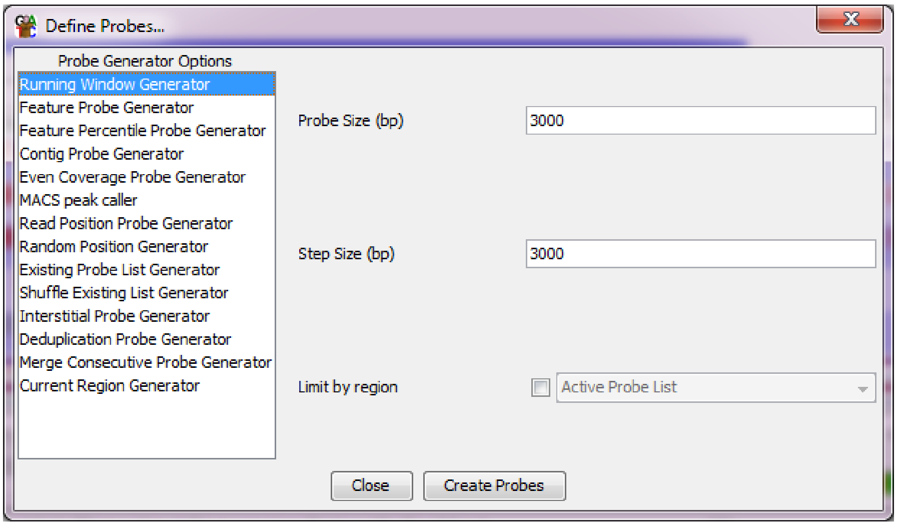
2. For every probe you calculate an associated value. This value derives in some way from the pattern of reads which occur within the probe region.

Once you have quantitated your data you will see the calculated values show up in your chromosome view, and you can start to use the filtering options.

Defining Probes

The first step in quantitation is defining your set of probes. You would normally only do this step once per analysis. SeqMonk only stores one set of probes, so if you create a second set of probes it will delete all of your existing probes and probe lists.

To start creating your probes select Data > Define Probes.


Define Probes...
Probe Generator Options
Running Window Generator
Feature Probe Generator
Feature Percentile Probe Generator
Contig Probe Generator
Even Coverage Probe Generator
MACS peak caller
Read Position Probe Generator
Random Position Generator
Existing Probe List Generator
Shuffle Existing List Generator
Interstitial Probe Generator
Deduplication Probe Generator
Merge Consecutive Probe Generator
Current Region Generator
Close
Create Probes

Probe Size (bp) 3000

Step Size (bp) 3000

Limit by region □ Active Probe List

There are several different strategies you can use to define where to put your probes, and these are listed on the left. As you click on the different entries on this list you will see that the options for that entry will be displayed on the right.

One generic option is to not create probes over a region which contains no data for any of your DataSets. This option will reduce the number of probes generated, but will mean that the probe generation will take longer to run.

Below is a list of the most commonly used probe generators, and a description of what they do. Descriptions for all of the other generators can be found in the programs help files.

Running Window Probe Generator

The simplest way to generate probes is to just make a regularly spaced set of probes along the genome. In SeqMonk you do this with the Running Window probe generator. All you need to specify is the size of window you want to use and the step size between the start points of adjacent windows. When you first open the options window SeqMonk will try to guess an appropriate window size for you, but you are free to change this.

If for example you wanted windows of 10kb, which had a 2kb overlap with the previous window you would set the window size to 10000 and the step size to 8000.

If you don't want the tiled probes to cover the whole genome then you can also restrict the generation to be constrained either within the region you're currently looking at, or to be within the probes which are currently defined in your project.

Feature Probe Generator

The feature probe generator allows you to design probes around a particular class of features. Once you have selected the feature type around which you want to design you can then choose to make a probe within the feature and upstream or downstream of the feature. SeqMonk will apply the appropriate correction if your feature is on the reverse strand.

For example, if you wanted one probe per gene you would select Gene as your feature class and then make a gene probe from 0 to 0 (no bases upstream or downstream – just the area covered by the feature). For a promoter probe you might select mRNA as the feature type and then make an upstream probe from -1000 to 0 (note that if two or more mRNAs had the same start point you would generate duplicate probes).

Read Position Probe Generator

In its simplest form this generator puts a probe over every different read position seen in your data. You wouldn't want to do this in most library types since you'd generate millions of probes (which will be slow), but if you have a restricted number of possible positions (eg for intron analysis, restriction based libraries or amplicon sequencing) then this can be useful. In other cases you can use the option to group different positions together, so that you can make probes of every 100 adjacent read positions for example. This can be a nice way to make probes with even coverage and this type of probe generation is commonly used in bisulphite sequencing.

MACS Peak Caller

If your data comes from a ChIP-Seq experiment you might want to make probes over regions which are significantly enriched for reads, a so called peak detection. SeqMonk uses the methodology employed by the MACS peak caller to do peak detection.

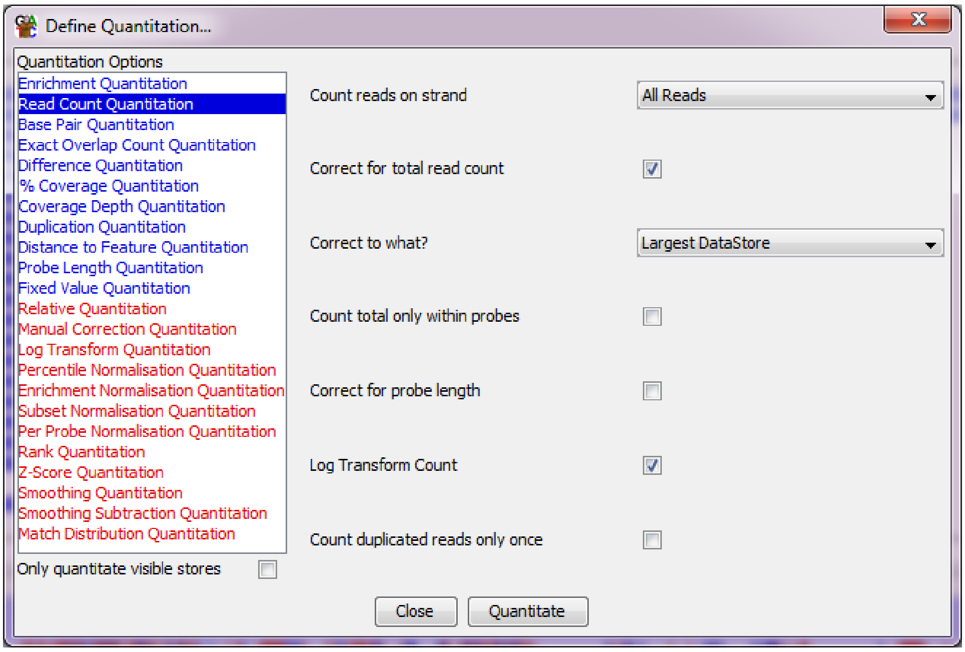
Current Region Generator

Sometimes you have a very specific question in mind and only need to make a single measurement per-sample. In this case you can use the Current Region Generator which will make a probe set consisting of a single probe covering the region you are currently looking at.

Quantitating Data

Once you have defined your probes you will need to quantitate your data. Again, there are several ways to do this depending on what you are trying to measure. In contrast to probe generation though it is likely that you will quantitate your data more than once in an analysis. As long as you keep your probe definitions the same you can requantitate your data whilst keeping all of your previous probe lists.

The probe quantitation options will automatically appear when you have finished generating your list of probes, and you can get to them manually at any time by selecting Data > Quantitation.



Quantitation Options
- Enrichment Quantitation
- Read Count Quantitation
- Base Pair Quantitation
- Exact Overlap Count Quantitation
- Difference Quantitation
- % Coverage Quantitation
- Coverage Depth Quantitation
- Duplication Quantitation
- Distance to Feature Quantitation
- Probe Length Quantitation
- Fixed Value Quantitation
- Relative Quantitation
- Manual Correction Quantitation
- Log Transform Quantitation
- Percentile Normalisation Quantitation
- Enrichment Normalisation Quantitation
- Subset Normalisation Quantitation
- Per Probe Normalisation Quantitation
- Rank Quantitation
- Z-Score Quantitation
- Smoothing Quantitation
- Smoothing Subtraction Quantitation
- Match Distribution Quantitation

Only quantitate visible stores □

Count reads on strand: All Reads

Correct for total read count ☑

Correct to what?: Largest DataStore

Count total only within probes □

Correct for probe length □

Log Transform Count ☑

Count duplicated reads only once □

As with the probe generation there are a selection of different ways to quantitate your data listed on the left and as you select each one their options will appear on the right. Some quantitations can only be applied on top of an existing quantitation - these appear in red in the list on the left, and will not be visible when you initially quantitate your raw data.

Once your quantitation is complete, a record of the options you used will be present in the All Probes list, accessible from your data view. If you double click on this you can see the options used for both probe generation and quantitation.

Read Count Quantitation

The simplest kind of quantitation is to count up the reads which overlap with each probe. Depending on how you intend to compare your datasets you may want to log transform this count. You can select to count reads on the forward or reverse strands only, but the default is to use reads on either strand.

You can also apply some corrections to your counts. If you have more than one dataset and you would like to perform a comparison then the counts in each set will be affected by the total number of reads in the dataset, and this value can vary greatly between runs. If you choose to correct for total read count, then the count will be adjusted relative to the dataset with the highest number of reads. This means that you can compare the counts between different datasets with different total read counts.

Depending on how you generated your probes you may also find that your probes are different lengths. When performing a simple count this will tend to overemphasise the importance of longer probes (since they will naturally produce a higher count). The corrected counts therefore give you a way to remove these two influences. If you choose to correct for probe length then the count for each probe will be corrected by a scaling factor (1000 / probe length), so that you can compare the values for probes of different lengths in the same dataset.

You should be aware that using these correction options can give misleading values where the absolute count for a probe is very low. If you have a very short probe with one reading in it this can produce an artificially high corrected value. It is therefore normally a good idea to combine a corrected count with a filter for a minimum number of reads per probe so you see only the reliable data. This will be discussed further when we get to filtering.

Base Pair Quantitation

This is very similar to the read count quantitation and has all the same options, the difference is that instead of just counting the number of reads which overlap each probe the base pair quantitation counts how many bases from each read overlaps. If you have many reads only partially overlapping a probe (commonly seen in mRNA-Seq for example) then this method will provide more relevant quantitation.

Enrichment Quantitation

This is the simplest quantitation method and has no options at all. It quantitates every probe with the relative enrichment of the data in that region compared to a completely even distribution of the same data over the whole genome. The enrichment is expressed as a $\log_2$ ratio relative to the even distribution so a value of 1 would indicate a 2-fold enrichment.

Difference Quantitation

In some datasets the relative proportion of forward and reverse reads may be relevant. In these cases, you can use a difference quantitation to compare the two strands. You can choose whether you want to do a subtraction of the count from one strand from the other, you can divide one by the other or you can calculate a percentage of one count relative to the other. For the division you can choose whether to log transform the ratio so as to always put difference on a linear scale. It can be useful to use the strand information to convey some other property for each read for example you could make forward be an exact match to the template and reverse be a read containing one or more SNPs. If you do this, then you could use the difference quantitation to pull out regions of high SNP density.

Percentage Coverage Quantitation

Another useful measure in some instances is to know what proportion of your reference sequence is covered by reads in the current data set. The percentage coverage will calculate this value for you. It doesn't matter how great the depth of coverage is over the parts which are covered, all it works out is what proportion of each probe is covered.

Running the Quantitation

When you have selected your options you can press 'Quantitate' to start the quantitation. Depending on the options you have set and the number of probes you defined the quantitation can take up to a couple of minutes to complete.

To speed up the quantitation process there is a checkbox at the bottom left of the quantitation options which allows you to only quantitate the currently visible data stores.

Viewing Quantitated Data

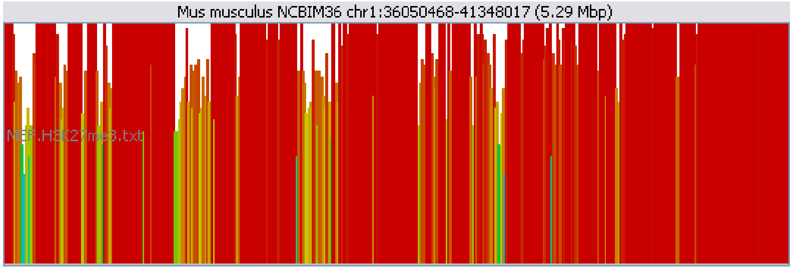
As soon as you have performed a quantitation the data tracks in the chromosome view will be updated to show the new quantitations. Once you can see this data there are a couple of things you may well need to change to make sense of it.

Changing the Data Zoom level

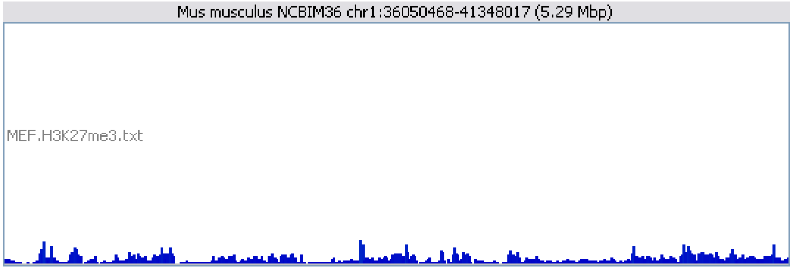
When you do a new quantitation by default the data zoom level will be adjusted to suit the data range of the currently visible data stores. You can also set the level of zoom manually. The window which opens also provides a scale bar for the current view which you can later export if you want to use it for figures.

To adjust the data zoom level, you should select View > Set Data Zoom Level (or just press Ctrl+Z). You can then drag the slider up and down to change the zoom level to something appropriate for your data.

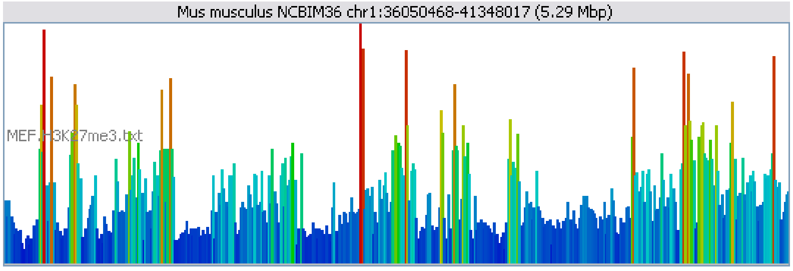
The table below shows the same region of quantitated data at three different zoom levels so you can see the effect the data zoom can have.



Too High



Too Low

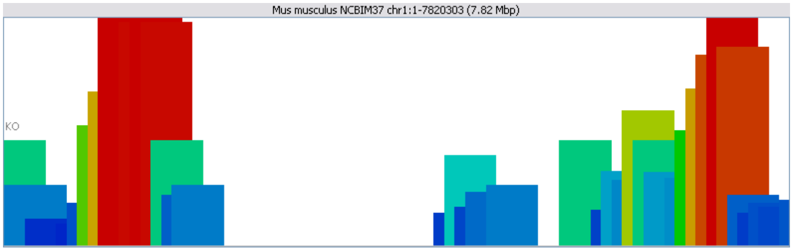


Just Right!

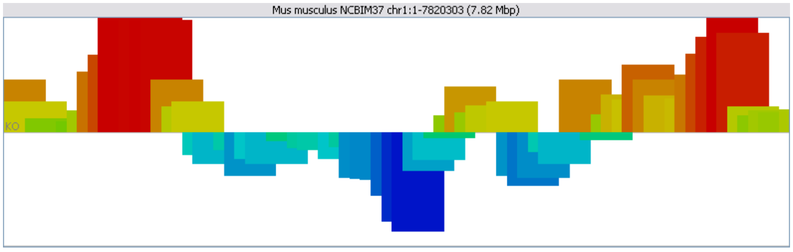
Changing the Scale type

After a quantitation SeqMonk will try to guess whether a positive, or positive and negative scale is more appropriate for your data and will adjust the view accordingly. If you use a positive and negative scale, then the scale is always symmetrical around the origin. If your data should have used a positive and negative scale, but SeqMonk incorrectly assigned a positive only scale to it then you can change this manually using the "Quantitated data scale" option in the display preferences.

If you're using log transformed data then you may find that all of your empty probes end up with negative values, whilst your probes containing data are positive. In this case it's often clearer to show only the positive values. This can have a dramatic effect on how you interpret your data so make sure you've turned this on if you need it.



Positive



Positive and Negative

Decluttering the Display

The default display for a data track shows both the raw reads and the quantitated data. In many cases though this is too much information. Once you have quantitated your data you may want to remove the raw reads, since these will clutter the display and slow down the speed at which the display can update. If you don't intend quantitating your data, you may want to show just the raw reads and give them all the available space in the data track.

You can make these changes by altering the “Display which data” option in the display preferences. Choices you can make are:

- Both Reads and Probes
- Just Reads
- Just Probes

Further Display Options

More options for changing the display of data in the chromosome view can be found under the display preferences. These include options for changing the type of graph used to represent the quantitated data, and the colour scheme used for the quantitation graphs.

For very large numbers of tracks it might, for example, be advisable to turn off the display of raw reads and to represent the quantitation as coloured blocks. You can play around with the options available to see what works for your particular data type.

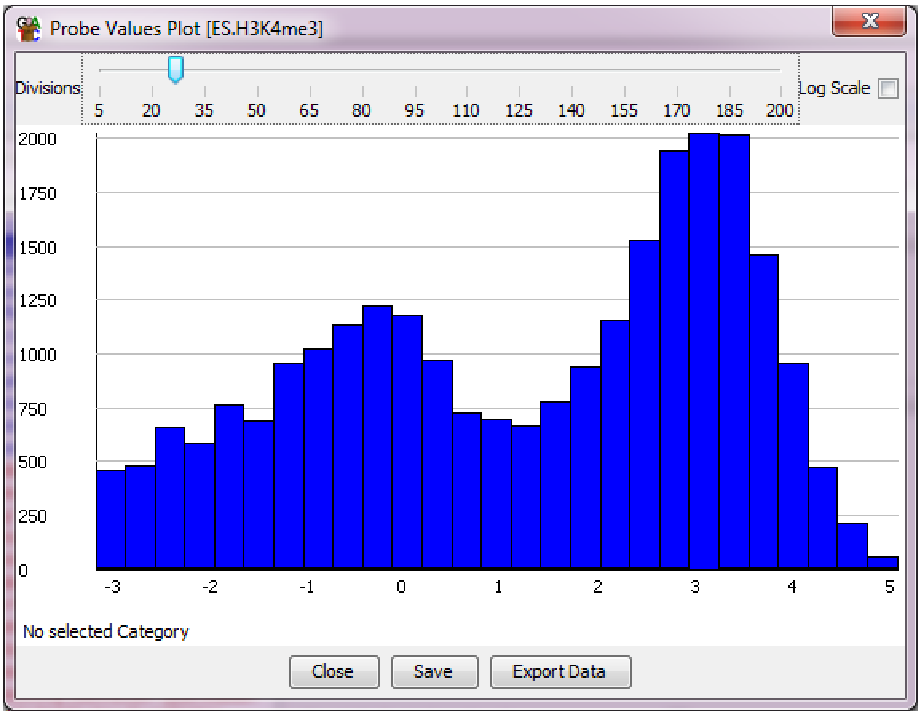
Quantitated Data Figures

Once you have some quantitated data there are a whole range of figures available to you under the Plots menu to help you interpret your data. We'll show a couple of example figures you can create which might help you interpret your data.

In general, unless a plot specifically says otherwise, the data shown in a plot will represent whatever values were generated by the last quantitation you ran. If you can't remember how you quantitated your data then you can double click on the “All Probes” list to see. Plots to display a single data store will plot whichever store you currently have selected in the Data View (and will generally refuse to plot if you don't have anything selected). If plots can plot multiple stores then they will plot whichever stores are visible in your chromosome view. To change the data which is plotted, just change the tracks which you have visible.

Distributions - The Probe Value Histogram

If you right click on a DataStore in the Data view you can select "Probe Value Histogram" from the popup menu which appears. This is a good way to let you see the range of values found in the probes in one of your data stores.

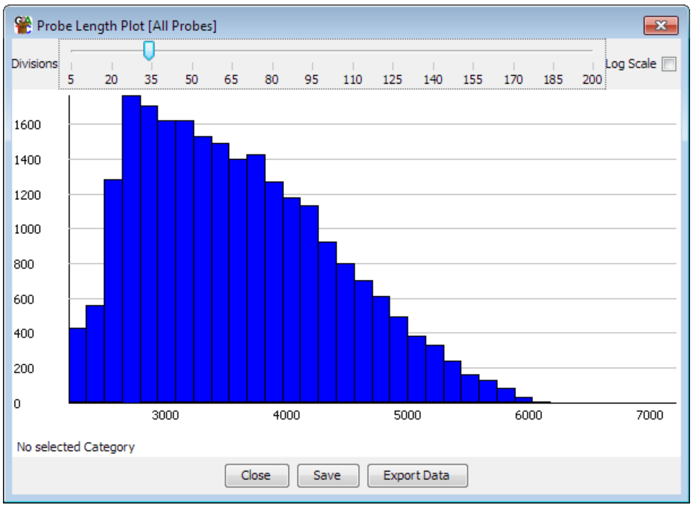


You can use the slider at the top to change the number of bins in your histogram, and if you have a large range of counts you can choose to plot the histogram on a log scale. If you put your mouse over one of the bins it will turn red, and the status bar at the bottom will tell you which range of values and how many probes were found in that bin. You can zoom into part of the plot by dragging the mouse within the plot window. You can zoom back out by right-clicking.

The probe value histogram operates on the currently active probe list, so if you have a probe list selected when you construct a histogram, only probes from within that list will be used to generate the plot.

Distributions - The Probe Length Histogram

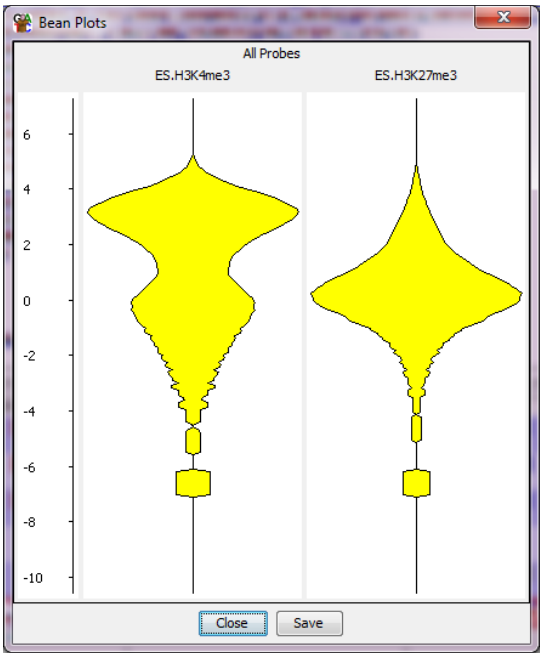
If you right click on a ProbeList in the Data view you can select "Show Probe Length Histogram" from the popup menu which appears. This is a quick way to see how long the probes in your list are.



The probe length histogram operates on the currently active probe list, so if you have a probe list selected when you construct a histogram, only probes from within that list will be used to generate the plot.

The Bean Plot

If you want to compare the distribution of probe values between several different DataStores you can use the Bean Plot. This provides a summarised view of your distribution of values and aligns plots from several different DataStores so you can quickly see any differences.



You can create a Bean plot by selecting Plots > Bean Plot > Visible Data Stores. You can choose whether you plot shows the same probe list over the currently visible data stores, or you can select multiple probe lists to display for the currently active data store.

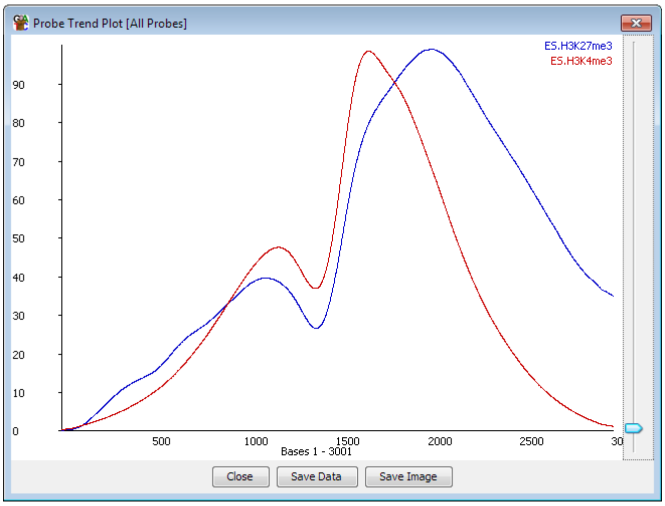
If you have data which is pretty normally distributed then you could use the BoxWhisker plot instead of a bean plot. If you only wanted to compare the mean values without showing the full distribution then you could try the Star Wars (confidence interval) plot instead.

The Probe Trend Plot

The probe trend plot is a way of looking for generic pattern which happen across all probes in a set of probes. It's different to other plots in that it doesn't use the pre-quantitated data for each probe, but makes its own calculation within each probe and then creates a profile for a 'generic' probe.

For example, if you had designed probes over promoters then you could see if there was a change in the distribution of reads across the promoter.

The probe trend plot will show a graph for every visible data store when it is launched, and will use the currently selected probe list. If all the probes in your set are the same size, then it will generate a plot at 1bp resolution. If they are of different sizes, then it will generate a relative plot where you can look at the distribution in percentiles across each probe.



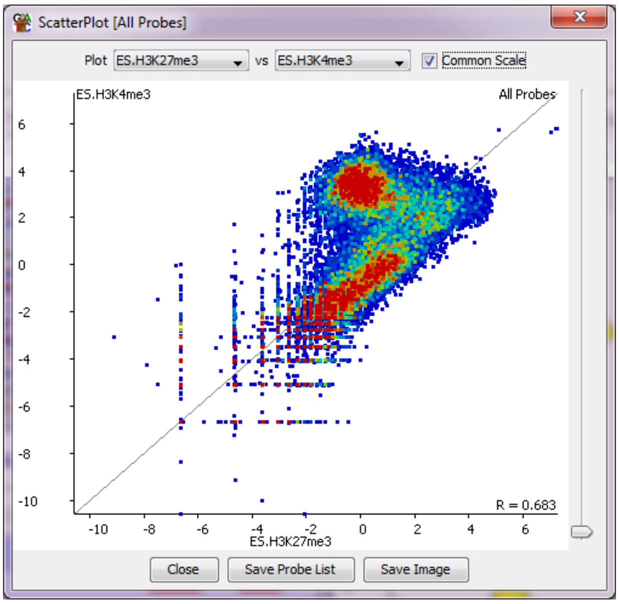
After calculating the plot, you can apply smoothing to the data by using the slider on the right of the plot. You can create a probe trend plot by selecting View > Probe Trend Plot.

For a more detailed view of the same type of data you could try the Aligned Probes Plot which represents the read density as colour depth in a histogram.

For displaying trends in existing quantitated values over larger features (eg if you have running window quantitation and want to summarise this per-gene) then you could use the Quantitation Trend Plot.

The Scatterplot

You can use a scatterplot to directly compare the quantitated values from two data stores. Simply select View > Scatterplot and the select the stores you want to plot from the drop down boxes at the top of the window.



You can use the slider on the right of the plot to change the size of the points plotted. You can also move your mouse around the plot to update the text at the bottom to show you the quantitation at the current position. The bottom right of the plot shows a Pearson's correlation R value for the currently plotted data.

If you drag your mouse within the plot you can select a group of points. The selection is done based on the difference between the two datasets and you can use the "Save Probe List" option at the bottom to save the list of selected points.

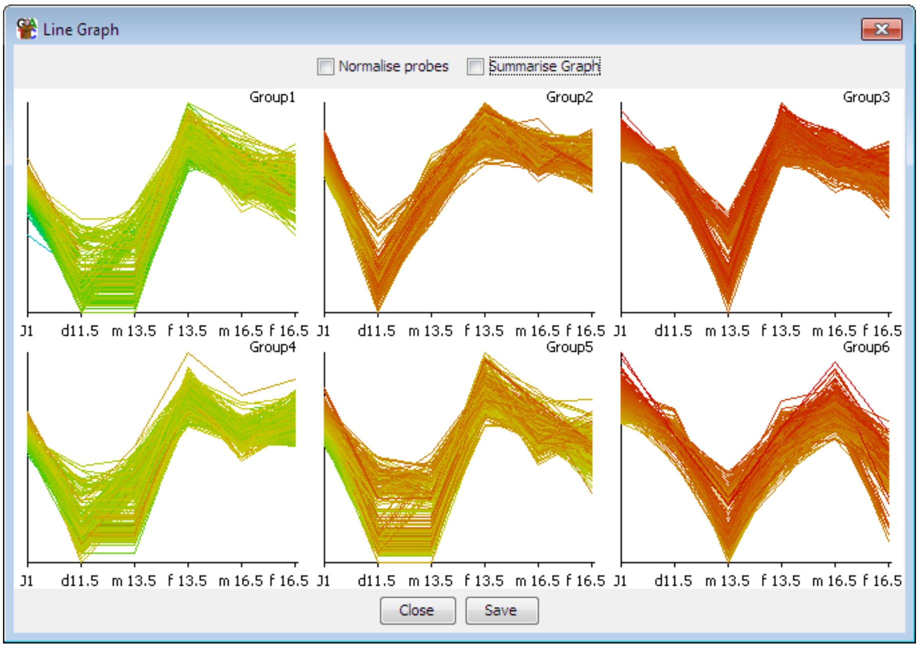
One nice options it that once you have multiple probe lists available in your project then you can plot out the points for one list, and then use the Highlight Sublists button to differently colour another sub-list. This can be really nice to allow you to see the results of filtering you've done for example.

An alternate, related, graph is the MA plot. This simply displays the difference between the two datasets plotted against the mean quantitated value. It's pretty much a scatter plot rotated 45 degrees to the right, but it's quite nice as it makes more efficient use of the available plotting space.

Other similar plots are the variation plot which works on replicate sets and plots the variance (there are a few different variance measures available), against the mean quantitation in the replicate set. It can be used to identify unusually variable (or consistent) probes within a set. There is also the volcano plot which we will describe in more detail later, which can be used to review the results of a statistical analysis. This plots the log p-value from a test against the absolute difference, allowing you to judge how strong a candidate in an analysis actually is.

The Line Graph

The line graph is a simple way to show the relative levels of quantitation for a set of probes across a number of different data stores. It will display the currently visible data stores and you can choose to show just the active probe list, or select a set of lists to display simultaneously. Inside the plot you have options for emphasising differences through normalisation, or showing just a summary of the probes rather than each individual value.


Line Graph
□ Normalise probes
□ Summarise Graph
Group1
Group2
Group3
Group4
Group5
Group6
J1
d11.5
m 13.5
f 13.5
m 16.5
f 16.5
Close
Save

Requantitating Data

Requantitating data is a common occurrence in a SeqMonk analysis. It's often useful to quantitate your data several different ways and filtering each quantitation. Changing the quantitation options will not affect your probe lists and the filtered sets you previously generated will still be present after requantitating.

Generating a new Probe Set

If you choose to rerun probe set generation after having quantitated your data, then the new probe set you produce will replace your existing probe set. This means that any filtered lists you had generated for the old set will also be lost.

Since it can be useful to look at the same dataset using a number of different probe sets the easiest way to achieve this is to save multiple projects, one for each probe set.

Another useful trick is that if you are going to replace a probe set, but want to keep some record of an interesting set of regions you had filtered then you can do this by turning them into an annotation track. To do this simply right click on a probe list and select "Convert to annotation track". The currently selected probe list will then be turned into an annotation set and will appear in the Annotation Sets folder in the data view. These annotations will still be present after the original probe set is replaced.

Filtering your Data

Once you have quantitated your data you can then begin to use the various filtering tools to try to identify regions of interest. These can all be found under the Filter menu.

The filtering tools allow you to generate Probe lists. A probe list is just a subset of the full complement of probes. Most filters will start from the currently selected Probe list, so that you can apply sequential filtering – using several different criteria to identify your final probes of interest.

As you start to run filters you will see a tree of Probe Lists building up under your All Probes list in your Data View. Generating a new probe set will create a new All Probes list and will therefore lose all of the probe lists under it.

Applying Filters

In general, there are two types of filter – those which filter based on the quantitated data, and those which filter based on probe or genome properties (how big probes are, which chromosome they're on etc etc). Most filters will analyse each probe they test individually, although a few operate in windows, in which case all probes within a window pass or fail as a set.

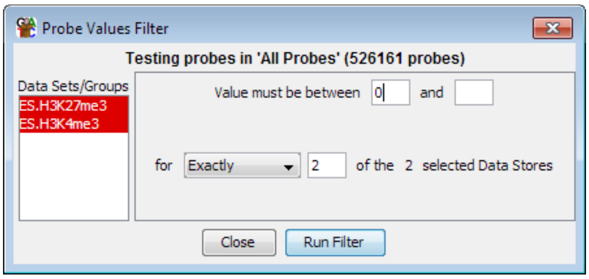
Every probe list will be given an automatic name, but you can change this either when it is created or at any subsequent point. All probe lists appear in a tree structure under the Probe List section of the Data view. The structure represents the order in which the lists were created, and the parent / child relationships between them.

If you select a probe list in the data view then only probes within the list will be shown in the chromosome view, and subsequent filters and plots will usually operate starting with the currently selected probe list.

Below is a list of some of the most common various filtering options.

Filter on Value

This is the simplest and most useful filter. It allows you to specify upper and lower limits for the quantitated value for probes you want to keep. You only have to specify one of the two limits (you can of course provide both if you want to). You must also then select a number of data stores and you can specify in how many of these the filter must be passed for the probe to be included in the filtered list.


Probe Values Filter

Testing probes in 'All Probes' (526161 probes)

Data Sets/Groups
ES.H3K27me3
ES.H3K4me3

Value must be between 0 and

for Exactly 2 of the 2 selected Data Stores

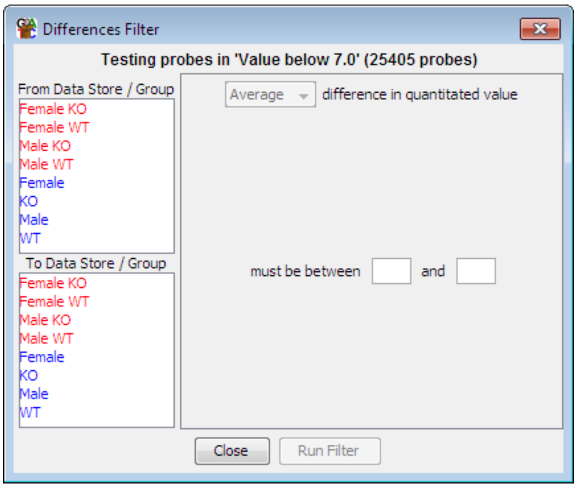
In this example both of the female samples must have a value greater than or equal to zero for the probe to pass the filter. Because the upper limit was not specified then there is no limit to how high the value can be.

The values used for this filter are whatever the current quantitated values on your probes are. Since these can change over time (if you rerun the quantitation tool), the results from this filter will show you the quantitation which was present at the time the filter was run. This is generally true for all filters which use quantitated values to make their selections.

A related filer is the Position in Distribution values filter. This is the same as the values filter except that instead of putting in absolute values you put in percentiles. This would let you find, for example, the highest 10% of probes.

Filter on Value Differences

In the value differences filter you are looking not at the absolute values for any data set but at the differences between datasets.



Differences Filter

Testing probes in 'Value below 7.0' (25405 probes)

Average difference in quantitated value

must be between and

The layout of the options looks very similar to the values filter. In this case though you must select at least one data set or group from each of the two lists on the left. You can then specify upper and lower limits for the difference value between the sets.

If you only select two data sets then the filtering is simple as you only have one difference value. If you select more than two data sets though the filter will work out all pairwise differences between the selected sets. You can then specify if you want to filter on the average, minimum or maximum difference from all of the pairs.

Filter by Position

Filter by position allows you to pull out the set of probes in a particular genomic region. You must specify a chromosome and optionally a start and end position. When you open this filter it will show you the region you are currently viewing in the chromosome view, so if you want a quick way to capture the probes in the current view then you can do this with the position filter.

Filter by Probe Length

This is a simple filter which allows you to select a size range for the probes you want to keep.

Filter by Feature

The feature filter allows you to specify probes according to their relationship to certain feature types. Previously we have seen that we can generate probes around features, but this filter allows us to filter probes generated by any method according to their proximity to certain features.

You can select any loaded feature class as the basis for the filter and there are a range of options to specify how close to this feature type and in what direction your probe needs to be to pass the filter.

Combining Probe Lists

There are a small set of filters which can be used to let you combine existing probe lists together. The different tools are good for different types of combination.

The **Logically Combine Two Lists** filter lets you do a binary logical combination of exactly two lists. The relationship between them can be AND, OR or BUTNOT.

The **Intersect Lists** filter lets you select as many lists as you like, and for each one you can specify whether the passing probes should be present or absent in that list (or whether it should be ignored). Using this you can build up complex logic across multiple lists.

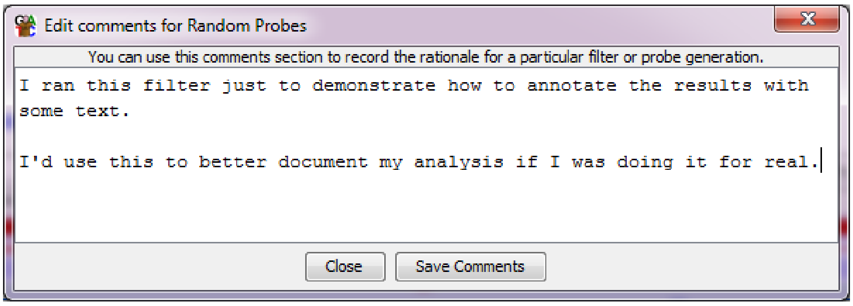
Where you need more flexibility you can use the **Collate Multiple Lists filter**. This lets you select several lists and then say in how many of them a probe must be present in order to select it.

Statistical Filters

There are a large number of statistical filters which are suitable for different kinds of data and experiment types. We will cover these in a later section. In general though they operate in the same way as any other filter, taking in a starting list of probes and producing a subset of that list.

Filter Comments

To make a better record of your thought processes when running an analysis in SeqMonk you can add your own notes to probe lists so that you can remember why you applied the filter in the way you did. Each filter will record the settings used when it was run, but you can add your own comments to say why those settings were used.



You can use this comments section to record the rationale for a particular filter or probe generation.

```
I ran this filter just to demonstrate how to annotate the results with
some text.

I'd use this to better document my analysis if I was doing it for real.
```

To edit the comments for any probe set simply right-click on the set in the data view and select "Edit Comments". You can view the automated and manual comments for a list by double-clicking on it (or right clicking and selecting "View List"), and these comments will also show up in the Probe List Description report.

Filter Annotations

As well as sub-setting a list of probes, a filter can also provide quantitative annotations to the resulting probe list. Depending on the nature of the type of filtering being performed a filter may choose not to provide any additional annotation (a genome position filter for example doesn't have anything useful to add), or it might add several values (a statistical filter will often add a p-value, false discovery rate and a difference value). These annotation can be seen by double clicking on any list, and will show up later when you create reports.

Exporting Data

There are a number of different ways to get information out of SeqMonk.

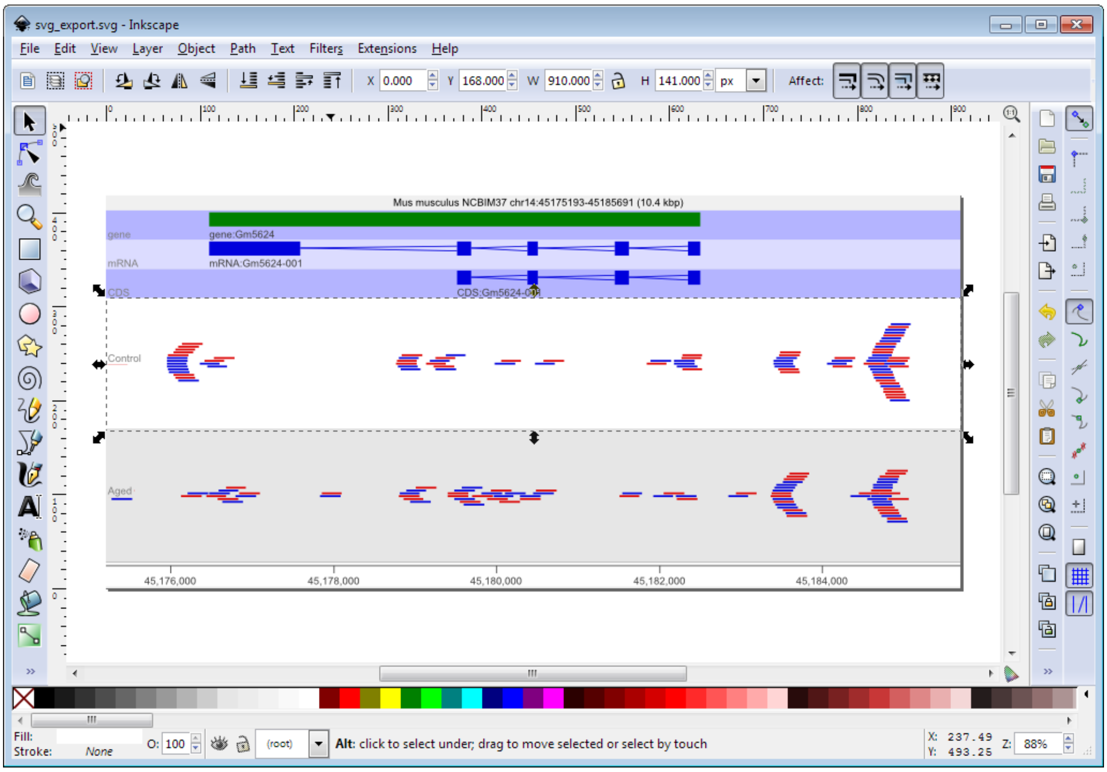
Exporting Graphics

If you have a chromosome or genome view which you'd like to export to a file, then you can do this at any point by selecting File > Export Current View. If you generate a plot, tree or histogram then all of these will include an option to save themselves as a file. Any plot which also performs quantitation or summarisation of the data will also include an option to save the data it calculated.

All graphics can be saved in one of two different formats, and the format you choose will depend on what you want to do with the figure.

- PNG Files are bitmap files. They are the equivalent of taking a screenshot, and the image you get will be exactly the same as you see on the screen. PNG files can be quickly inserted into most applications which handle images, but cannot easily be edited. They are generally not of sufficient resolution to use for publication.

- SVG Files are a vector image format. This means that you can easily edit them to show exactly what you want. You can also then create very high resolution images for publication. SVG files are not as widely supported as PNG so you may not be able to insert them directly into documents, but you can use programs such as Inkscape or Adobe Illustrator to edit them and make high quality png files from them.



X 0.000 Y 168.000 W 910.000 H 141.000 px

Fill: Stroke: None O: 100 (root)

Alt: click to select under; drag to move selected or select by touch

X: 237.49 Y: 493.25 Z: 88%

svg_export.svg - Inkscape
File Edit View Layer Object Path Text Filters Extensions Help
Affect:
Mus musculus NCBIM37 chr14:45175193-45185691 (10.4 kbp)
gene
gene:Gm5624
mRNA
mRNA:Gm5624-001
CDS
CDS:Gm5624-001
Control
Aged
45,176,000
45,178,000
45,180,000
45,182,000
45,184,000

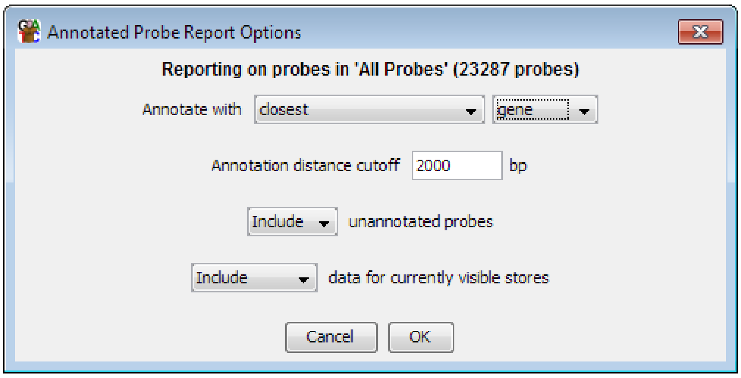
Creating Reports

If you have created a probe list for your regions of interest, you can just double click on it to view it, and then save that table to a file. However, this will only contain the probe names and positions. It may be more useful to compile this list into a report which includes not only the probe details but can also associate these with nearby features and can include the quantitated data for one or more data sets.

There are several types of report in SeqMonk, some based around probe lists and others providing more general information. Using these reports is generally the easiest way to export your analysed data from the program.

Annotated Probe Report

The annotated probe report can be created by selecting the probe list of interest and then going to Reports > Create Annotated Probe Report.



Reporting on probes in 'All Probes' (23287 probes)

Annotate with closest gene

Annotation distance cutoff 2000 bp

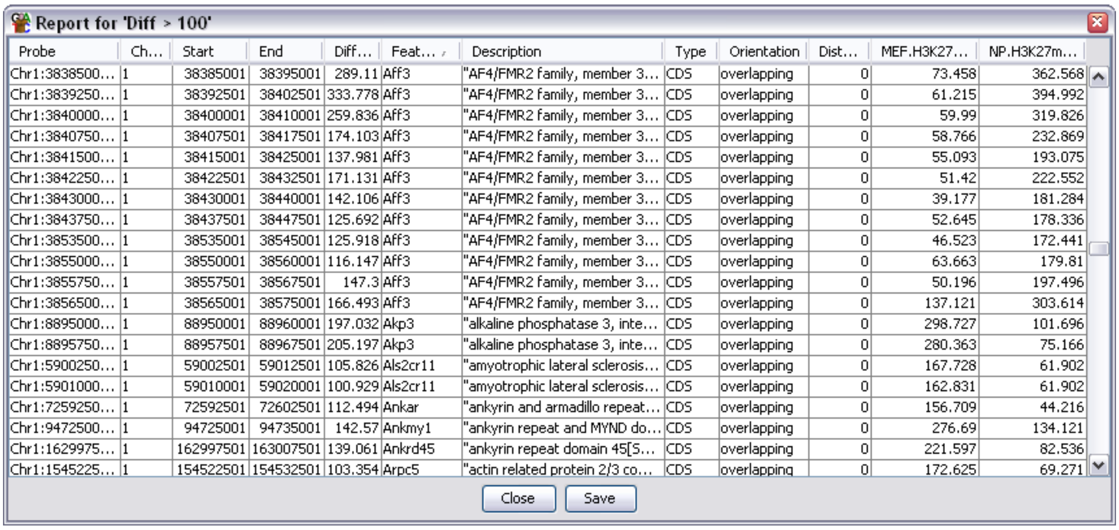
Include unannotated probes

Include data for currently visible stores

You can choose how you want to annotate each probe by selecting the type of feature you want to annotate with and then saying where the probe needs to be relative to the feature in order to become annotated. If more than one feature matches, then all of them will be included in the report.

You can choose whether to include probes which were not able to be annotated, and you can also choose whether you want to add the quantitated data for all of the currently visible data sets alongside the probe information.

When you create the report you will see something like this:



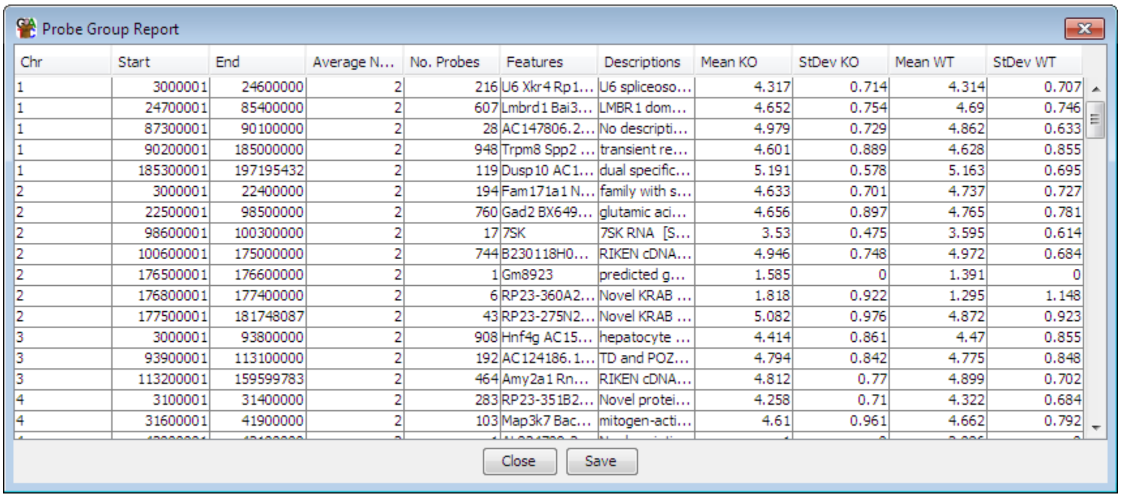
Probe	Ch…	Start	End	Diff…	Feat…	Description	Type	Orientation	Dist…	MEF.H3K27…	NP.H3K27m…
Chr1:3838500…	1	38385001	38395001	289.11	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	73.458	362.568
Chr1:3839250…	1	38392501	38402501	333.778	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	61.215	394.992
Chr1:3840000…	1	38400001	38410001	259.836	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	59.99	319.826
Chr1:3840750…	1	38407501	38417501	174.103	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	58.766	232.869
Chr1:3841500…	1	38415001	38425001	137.981	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	55.093	193.075
Chr1:3842250…	1	38422501	38432501	171.131	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	51.42	222.552
Chr1:3843000…	1	38430001	38440001	142.106	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	39.177	181.284
Chr1:3843750…	1	38437501	38447501	125.692	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	52.645	178.336
Chr1:3853500…	1	38535001	38545001	125.918	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	46.523	172.441
Chr1:3855000…	1	38550001	38560001	116.147	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	63.663	179.81
Chr1:3855750…	1	38557501	38567501	147.3	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	50.196	197.496
Chr1:3856500…	1	38565001	38575001	166.493	Aff3	"AF4/FMR2 family, member 3…	CDS	overlapping	0	137.121	303.614
Chr1:8895000…	1	88950001	88960001	197.032	Akp3	"alkaline phosphatase 3, inte…	CDS	overlapping	0	298.727	101.696
Chr1:8895750…	1	88957501	88967501	205.197	Akp3	"alkaline phosphatase 3, inte…	CDS	overlapping	0	280.363	75.166
Chr1:5900250…	1	59002501	59012501	105.826	Als2cr11	"amyotrophic lateral sclerosis…	CDS	overlapping	0	167.728	61.902
Chr1:5901000…	1	59010001	59020001	100.929	Als2cr11	"amyotrophic lateral sclerosis…	CDS	overlapping	0	162.831	61.902
Chr1:7259250…	1	72592501	72602501	112.494	Ankar	"ankyrin and armadillo repeat…	CDS	overlapping	0	156.709	44.216
Chr1:9472500…	1	94725001	94735001	142.57	Ankmy1	"ankyrin repeat and MYND do…	CDS	overlapping	0	276.69	134.121
Chr1:1629975…	1	162997501	163007501	139.061	Ankrd45	"ankyrin repeat domain 45[S…	CDS	overlapping	0	221.597	82.536
Chr1:1545225…	1	154522501	154532501	103.354	Arpc5	"actin related protein 2/3 co…	CDS	overlapping	0	172.625	69.271

If you want to see where any of these probes are you can simply double click on any line to jump to that point in the chromosome view. You can also press the save button at the bottom to export this information into a tab-delimited text file which you can open in a spreadsheet.

Probe Group Report

The probe group report is very similar to the annotated probe report except that you also get to specify a distance below which adjacent probes are considered to be part of the same feature and will be combined in the output. This gives a cleaner output where each of your regions of interest may contain multiple probes. The quantitation values are averaged over the set of probes which are grouped together and the mean and standard deviation of this set is shown in the final report.

As with the annotated probe report you can double click on any line of the probe group report to jump to that region in the chromosome view.



Chr	Start	End	Average N...	No. Probes	Features	Descriptions	Mean KO	StDev KO	Mean WT	StDev WT
1	3000001	24600000	2	216	U6 Xkr4 Rp1...	U6 spliceoso...	4.317	0.714	4.314	0.707
1	24700001	85400000	2	607	Lmbrd1 Bai3...	LMBR1 dom...	4.652	0.754	4.69	0.746
1	87300001	90100000	2	28	AC147806.2...	No descripti...	4.979	0.729	4.862	0.633
1	90200001	185000000	2	948	Trpm8 Spp2 ...	transient re...	4.601	0.889	4.628	0.855
1	185300001	197195432	2	119	Dusp10 AC1...	dual specific...	5.191	0.578	5.163	0.695
2	3000001	22400000	2	194	Fam171a1 N...	family with s...	4.633	0.701	4.737	0.727
2	22500001	98500000	2	760	Gad2 BX649...	glutamic aci...	4.656	0.897	4.765	0.781
2	98600001	100300000	2	17	7SK	7SK RNA [S...	3.53	0.475	3.595	0.614
2	100600001	175000000	2	744	B230118H0...	RIKEN cDNA...	4.946	0.748	4.972	0.684
2	176500001	176600000	2	1	Gm8923	predicted g...	1.585	0	1.391	0
2	176800001	177400000	2	6	RP23-360A2...	Novel KRAB ...	1.818	0.922	1.295	1.148
2	177500001	181748087	2	43	RP23-275N2...	Novel KRAB ...	5.082	0.976	4.872	0.923
3	3000001	93800000	2	908	Hnf4g AC15...	hepatocyte ...	4.414	0.861	4.47	0.855
3	93900001	113100000	2	192	AC124186.1...	TD and POZ...	4.794	0.842	4.775	0.848
3	113200001	159599783	2	464	Amy2a1 Rn...	RIKEN cDNA...	4.812	0.77	4.899	0.702
4	3100001	31400000	2	283	RP23-351B2...	Novel protei...	4.258	0.71	4.322	0.684
4	31600001	41900000	2	103	Map3k7 Bac...	mitogen-acti...	4.61	0.961	4.662	0.792

Both of the report formats give you the option to save the report to a file. The reports save as a tab-delimited text file. You should be able to open this directly in any spreadsheet application to allow you to do further filtering, plotting or annotation of your results.

Feature Report

The feature report is similar to the probe group report in that it allows you to collapse together the values from several probes. The difference is that instead of doing this based on proximity the individual probes are joined based on their overlap with members of a particular class of features. Each line in the report therefore represents one feature and all of the probes which overlap with it.

When assigning probes to features you can choose to include all overlapping probes, or only probes which overlap exactly with either the whole features, or one of its subcomponents (exons). This type of report is very useful when used in combination with the feature probe generator where you can design probes over every exon and then use this report to re-assemble the exons back to give an overall value for each transcript.

DataStore Summary Report

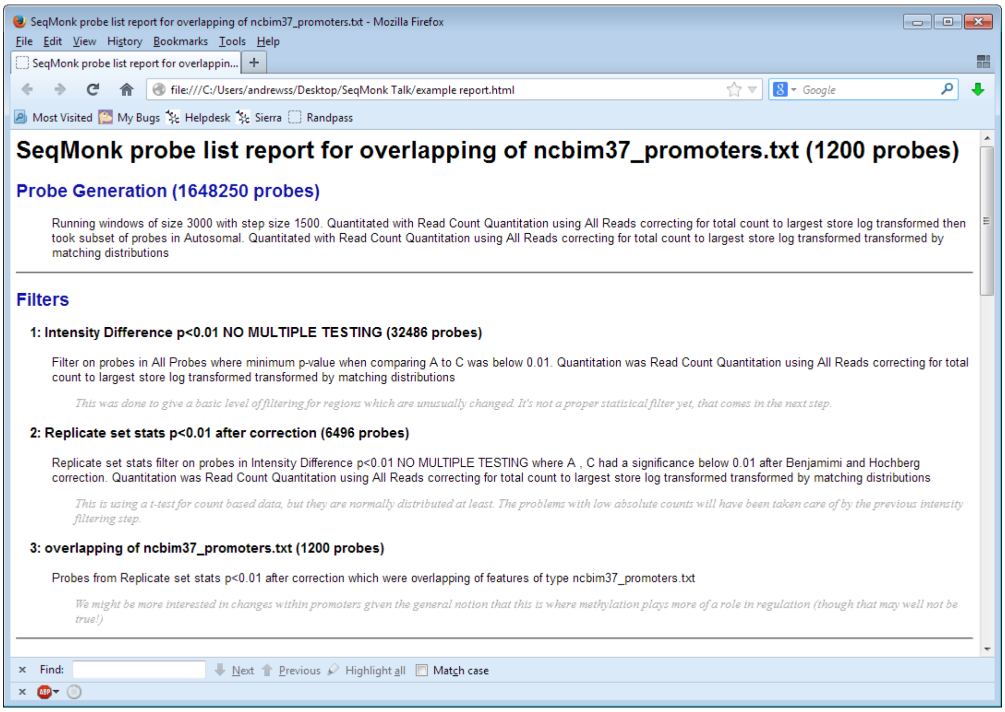
The DataStore summary report is a way of quickly collecting some simple statistics for all of the data stores you are currently looking at. In a tabular form it lists:

- The total / forward / reverse / unknown read count
- The mean read length
- The total read length

- The fold genome coverage (the total read length divided by the genome size)
- The total of all quantitated values
- The median quantitated value
- The mean quantitated value
- The number of valid quantitations. Some quantitation methods can generate null values where there isn't enough data to make an accurate quantitation. This value says how many probes were successfully quantitated.

Probe List Description Report

The probe list description report is a way to record exactly what you did to create a final list of candidate probes so that you can reproduce this analysis in future. It is an HTML report which collects the details of the probe generation, quantitation and filtering steps you undertook and combines this with details of the datasets you loaded so that you can get a complete picture of what you did.


SeqMonk probe list report for overlapping of ncbim37_promoters.txt - Mozilla Firefox
File Edit View History Bookmarks Tools Help
SeqMonk probe list report for overlappin...
file:///C:/Users/andrewss/Desktop/SeqMonk Talk/example report.html
Google
Most Visited My Bugs Helpdesk Sierra Randpass
Find: Next Previous Highlight all Match case

SeqMonk probe list report for overlapping of ncbim37_promoters.txt (1200 probes)

Probe Generation (1648250 probes)

Running windows of size 3000 with step size 1500. Quantitated with Read Count Quantitation using All Reads correcting for total count to largest store log transformed then took subset of probes in Autosomal. Quantitated with Read Count Quantitation using All Reads correcting for total count to largest store log transformed transformed by matching distributions

Filters

1: Intensity Difference p<0.01 NO MULTIPLE TESTING (32486 probes)

Filter on probes in All Probes where minimum p-value when comparing A to C was below 0.01. Quantitation was Read Count Quantitation using All Reads correcting for total count to largest store log transformed transformed by matching distributions

This was done to give a basic level of filtering for regions which are unusually changed. It's not a proper statisical filter yet, that comes in the next step.

2: Replicate set stats p<0.01 after correction (6496 probes)

Replicate set stats filter on probes in Intensity Difference p<0.01 NO MULTIPLE TESTING where A , C had a significance below 0.01 after Benjamimi and Hochberg correction. Quantitation was Read Count Quantitation using All Reads correcting for total count to largest store log transformed transformed by matching distributions

This is using a t-test for count based data, but they are normally distributed at least. The problems with low absolute counts will have been taken care of by the previous intensity filtering step.

3: overlapping of ncbim37_promoters.txt (1200 probes)

Probes from Replicate set stats p<0.01 after correction which were overlapping of features of type ncbim37_promoters.txt

We might be more interested in changes within promoters given the general notion that this is where methylation plays more of a role in regulation (though that may well not be true!)

Advanced Quantitation

Quantitation Pipelines

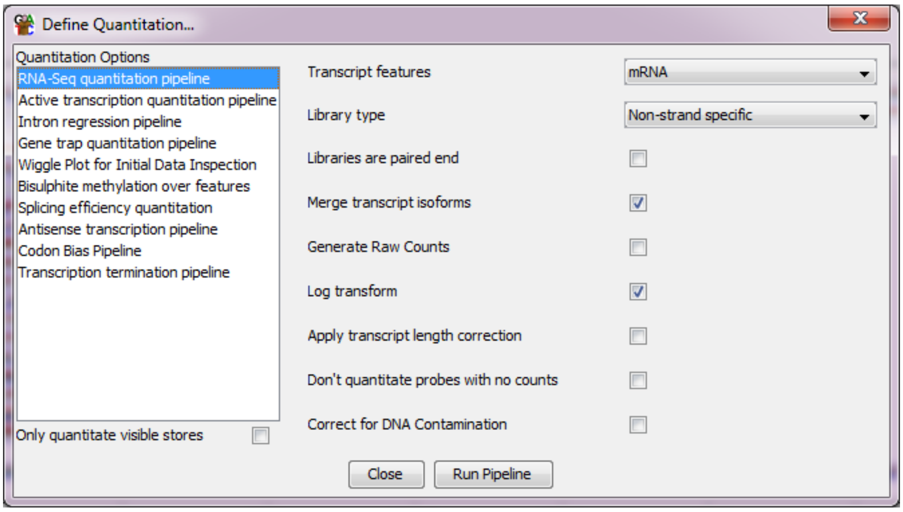
The traditional way to quantitate data in SeqMonk is to use a two-step process where the first step involves the creation of a set of probes over regions which will later be quantitated, and the second step assigns a value to each probe for each dataset based on the data.

In some cases, there are types of quantitation which can't be performed in this type of 2-step process, and there are also some common quantitations for which it would be convenient to have a simpler way to perform them. These use cases are what Quantitation Pipelines were designed for. They are automated quantitation processes which can perform probe generation, quantitation and potentially even filtering or other analyses in a single step. Quantitation pipelines can be accessed under Data > Quantitation Pipelines.

Some examples of the most commonly used pipelines would be:

RNA-Seq quantitation pipeline

The RNA-Seq pipeline uses a type of quantitation which can only be performed within a pipeline since it requires that quantitation and probe generation are performed simultaneously. The basic premise of the pipeline is that it is a read count quantitation over a set of features, but for multi-exon features it only counts the reads which sit over the exons of the feature and ignores those in the introns.



Define Quantitation...

Quantitation Options
- RNA-Seq quantitation pipeline
- Active transcription quantitation pipeline
- Intron regression pipeline
- Gene trap quantitation pipeline
- Wiggle Plot for Initial Data Inspection
- Bisulphite methylation over features
- Splicing efficiency quantitation
- Antisense transcription pipeline
- Codon Bias Pipeline
- Transcription termination pipeline

Only quantitate visible stores □

Transcript features: mRNA
Library type: Non-strand specific
Libraries are paired end □
Merge transcript isoforms ☑
Generate Raw Counts □
Log transform ☑
Apply transcript length correction □
Don't quantitate probes with no counts □
Correct for DNA Contamination □

The options you have within the pipeline allow you to choose which class of features you're going to use to quantitate your data. You can also choose the type of strand specificity in the libraries you're quantitating so you can only count the relevant reads.

The default quantitation from this pipeline is log2 RPM (reads per feature per million reads of library), but you can alter this. If you're going to use the quantitation in an external analysis tool such as DESeq or EdgeR then you will need uncorrected raw read counts, so there is an option to generate these. If you want to compare the expression levels of different genes within the same sample then you can also correct by read length to get log2 RPKM values, but this is not recommended when comparing expression values between samples.

By default, the RNA-Seq pipeline merges together the exons of different transcript isoforms for the same gene to give you a single per-gene expression value. You can choose to get output for each transcript isoform by unticking the 'merge transcript isoforms' box, but SeqMonk does not try to do a likelihood based assignment of reads to a single isoform, so reads which map against more than one isoform will be counted more than once in this mode. It also has an option to estimate and subtract signal from DNA contamination if that is a problem in your libraries.

Wiggle plot pipeline

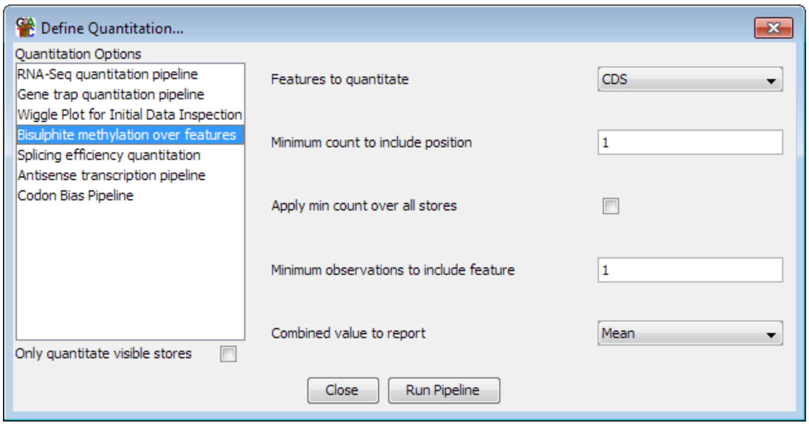
The wiggle plot pipeline is a convenience method, and everything in this pipeline could be performed in the normal quantitation options. This simply generates a set of running window probes and quantitates them with a corrected read count. This is an easy way to generate an initial unbiased quantitation of your data to look quantitatively at your read counts over a region of interest.

You can choose which region you want to analyse (just what you're looking at now, the current chromosome or the whole genome) and the pipeline will try to select a suitable window size for your probes.

Bisulphite methylation over feature

When analysing bisulphite data what is imported into SeqMonk are not the read positions, but the individual methylation calls. In a BS-Seq dataset each 'read' is only 1 base long and the strand of the read indicates the methylation state (forward = methylated, reverse = unmethylated).

You can calculate a percentage methylation value for a region of the genome using a standard difference quantitation by simply counting the total number of methylated and unmethylated calls within that region, however this has a number of problems. You could have very few calls which could result in a very unreliable overall methylation value. You could also have very biased coverage such that the majority of your reads come from one or two positions within the region which again might not produce a good overall average.


Define Quantitation...
Quantitation Options
RNA-Seq quantitation pipeline
Gene trap quantitation pipeline
Wiggle Plot for Initial Data Inspection
Bisulphite methylation over features
Splicing efficiency quantitation
Antisense transcription pipeline
Codon Bias Pipeline
Close
Run Pipeline

Features to quantitate: CDS

Minimum count to include position: 1

Apply min count over all stores ☐

Minimum observations to include feature: 1

Combined value to report: Mean

Only quantitate visible stores ☐

The bisulphite methylation pipeline provides a way to account for some of these problems. It provides a way to filter each call position by the degree of coverage and then produces an overall methylation value which weights each valid call position equally.

Quantitation Normalisation

SeqMonk provides a series of tools to quantitate your raw data. These include some normalisation options allowing you to correct for factors such as the total number of reads in your dataset and the length of the probe you're quantitating. Although this initial quantitation can provide useful data, in many cases it is susceptible to other sources of bias which might cause systematic differences between your datasets, and which might lead you to make incorrect conclusions about the differences between your data.

Common sources of bias might be:

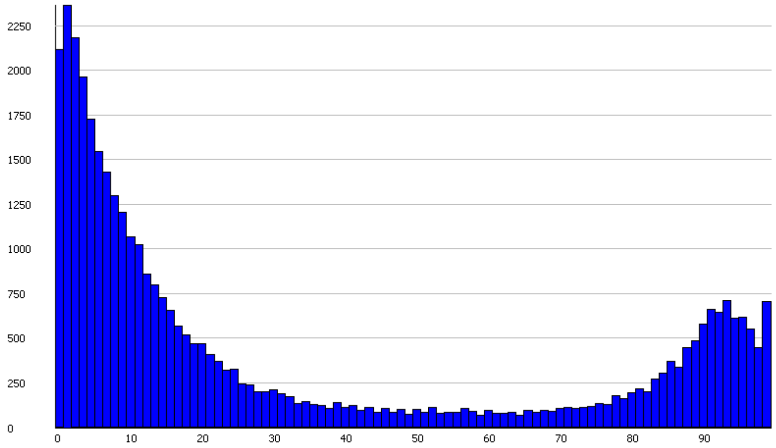
- Having greatly different numbers of reads between samples, meaning that low values may be measured with very different accuracies between data sets.

- Having different levels of PCR duplication between samples
- Having different read lengths
- Having different total amounts of signal between samples (e.g. RNA-Seq samples where there is more transcription in one sample than another.
- Having different degrees of mis-mapping contaminating sequences into different samples

Before you trust your quantitation you should therefore take some time to look at the set of quantitated values you have produced and compare these between your datasets. If there are systematic differences between your samples then you either need to think of why this might make sense biologically, or if you decide the differences are technical you can try to normalise the data so that their influence is removed.

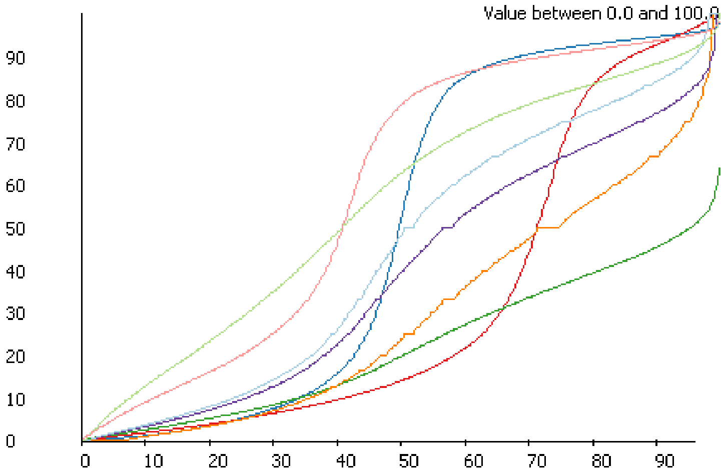
Plotting Quantitation Distributions

In order to assess whether there are issues with a set of quantitation which might necessitate performing additional normalisation there are a number of plots you could use. We saw some of these earlier where we used the probe value histogram or the Bean plot to look at distributions:



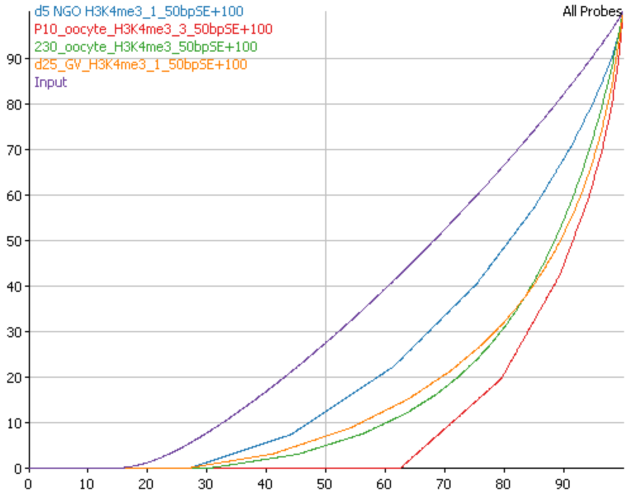
These plots can be useful in the assessment of distributions, but the easiest plot to compare distributions for normalisation is the Cumulative Distribution Plot. This plot simply shows you the path your quantitated data takes to get from the lowest value in your set to the highest. The x-axis is the percentile through the data, and the y-axis shows the current quantitation value for that percentile. Each line is a different data set.

Different datasets which show the same distribution of values (even if individual measures show large changes) should have virtually identical paths on this plot, and the aim of normalising data is to make the paths as similar as possible, or if they are too different to decide whether this reflects something of biological interest.



To spot more subtle differences in quantitation between datasets (particularly in ChIP-Seq data) you can also use a related plot called a QQ plot. This is another line graph type plot, but with different axes to the cumulative distribution plot.

In the QQ plot, for each data store the sum of the quantitations for all probes is calculated. The probes are then ordered from lowest to highest quantitation. The probes are then plotted sequentially with their percentile position (on the x axis) plotted againt the percentile quantitation of the summed quantitation seen to that point, plotted on the y axis. Data which is completely uniformly distributed will sit on a diagonal, and any enrichment bias will cause a curve below the centre line.



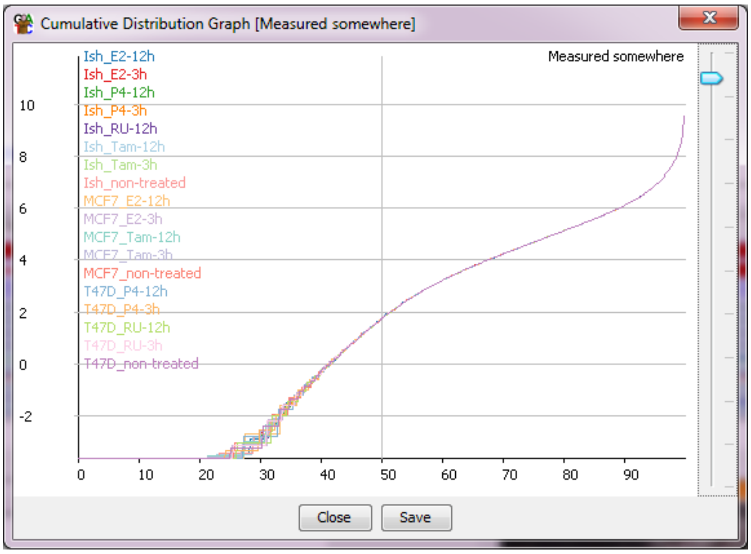
Applying additional normalisation

If these plots show that there is a global difference in the distributions of different samples then you may want to apply some additional normalisation or correction. There are a number of methods to do this available under Data > Quantitate existing probes. The methods shown in red are the ways of modifying the existing quantitations.

Match distribution quantitation

The simplest, and crudest, method of post-quantitation normalisation is to force all of the datasets to follow exactly the same distribution. Ideally you should use a more targeted method of normalisation after understanding the exact reason for the differences you observe, but in many cases it can be quicker and easier to just force a common distribution.

The match distribution normalisation works by generating an averaged distribution from all samples and then moving the quantitation of each probe in each sample to the closest position on the averaged distribution. The way the normalisation works, we guarantee that probes in a single dataset which started off with the same value will end up with the same value, and that the order of the quantitations in a data set will never change, but beyond that the normalisation can be completely non-linear.

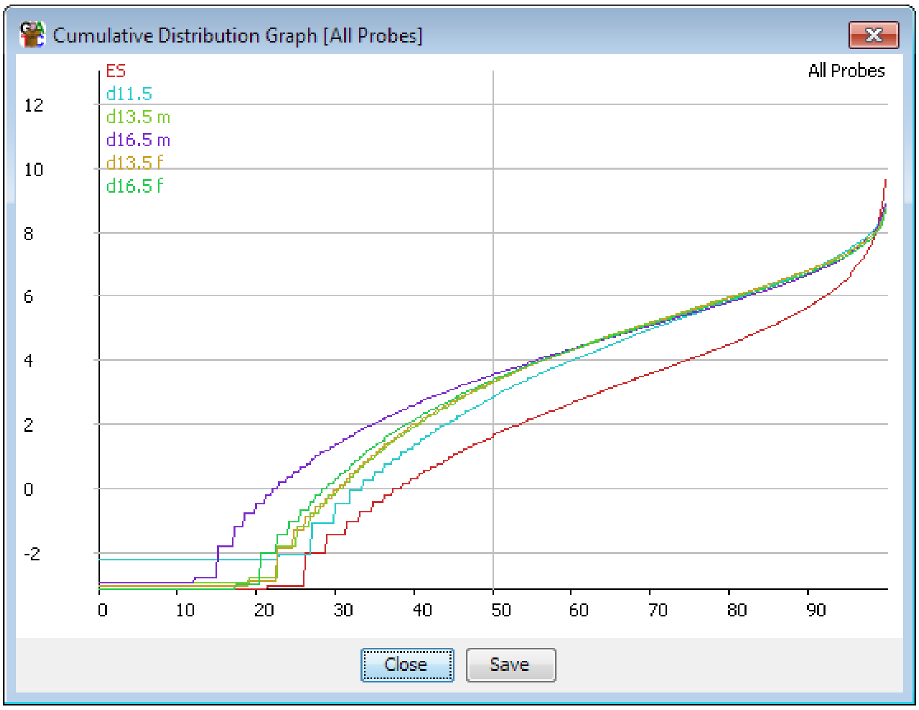


Distributions following this normalisation will always look really nice, but that's because you forced them too so don't get too carried away.

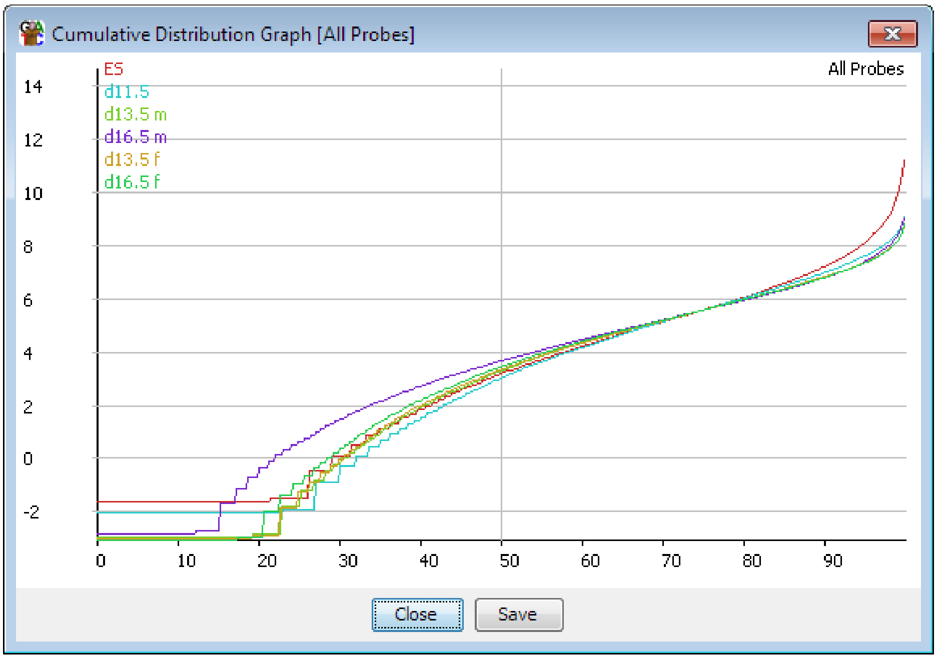
Percentile normalisation quantitation

A commonly observed pattern is that your distributions follow a similar path, but on a somewhat different scale. Thus the original quantitation would suggest that there is a consistent difference between your samples, and scatterplots would show an off-diagonal consistent relationship between the samples. This kind of discrepancy is usually the result of a failure of standard global normalisation, normally due to the presence or absence of a small number of very high coverage outliers (rRNA or Mitochondrial sequences are the most common ones, but there are others). Since the differences observed do not in most cases represent an actual biological change then it is reasonable to aim to normalise away this difference.

The simplest tool to achieve this type of normalisation is the Percentile Normalisation Quantitation. This quantitation method allows you to set a reference percentile in your data and the existing quantitations will have a correction factor applied to them such that their values at your chosen percentile will match exactly. Normally it makes sense to set the reference percentile to somewhere around 75% since this will be in the well measured portion of your data, but before any big changes which might occur at the extreme end of the distribution.



You can also choose what kind of correction will be applied. The default is an additive correction where a constant value is added to each point to get the distributions to match. The downside to this method is that the same correction will be applied everywhere, and will end up with different values being set for the empty probes in your different data sets. A more appropriate correction in these cases is to use a multiplying factor to correct your data. This will 'stretch' your distribution to match at the specified percentile and may more closely match the overall distributions, especially at the low end.



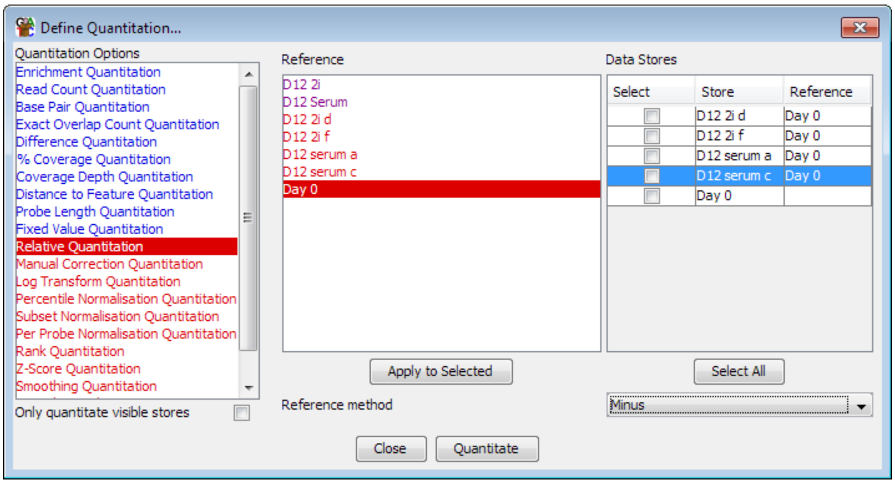
Since the same correction is applied to all points on your distribution this is a fairly safe and uncontentious correction to apply, but in some cases this correction alone will not be enough to completely match the distributions of your samples.

Enrichment normalisation quantitation

The enrichment normalisation quantitation is an extension of the percentile normalisation quantitation. It is used in cases where there is a non-linear difference between the values in different data stores. This most often happens when you get ChIP data with different enrichment efficiencies. In the enrichment normalisation instead of picking a single percentile to normalise to, you pick two. The data is translated (added) so that the lower values all match, and then multiplied to make the higher values match.

Relative quantitation

In some cases, you might want to normalise your samples against each other - quantitating sample A as the difference between sample A and sample B for example. There is a quantitation module which lets you apply this type of more complex normalisation called the relative quantitation module.



Select	Store	Reference
□	D12 2i d	Day 0
□	D12 2i f	Day 0
□	D12 serum a	Day 0
□	D12 serum c	Day 0
□	Day 0	

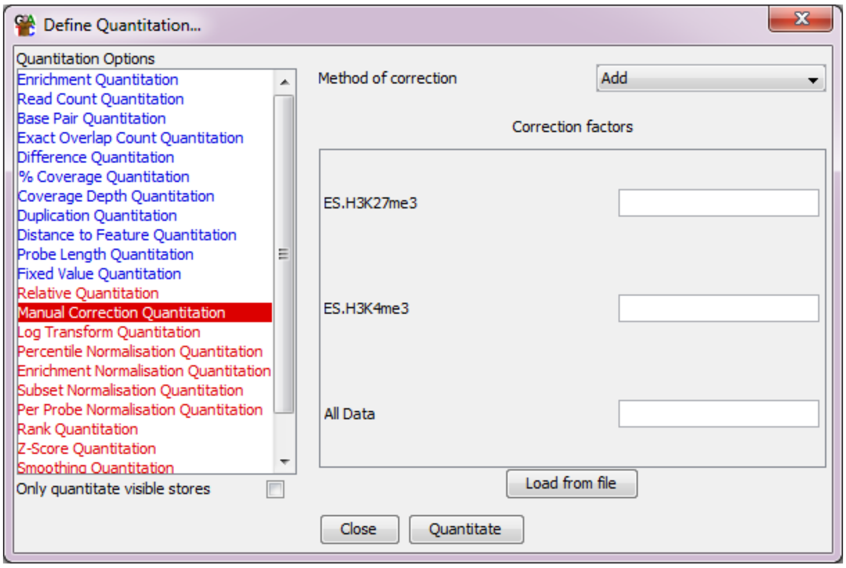
Reference method: Minus

This module allows you to pair up your samples and then choose what operation to apply between each sample and its selected reference. To match a sample and reference you select the reference from the reference list and then check the boxes next to the samples to which you want to apply this reference. Pressing the “Apply to selected” button will then assign that reference to those samples. Once you’ve paired up all of the samples you want to correct you can choose how you want to normalise (subtract reference, divide by reference etc.) from the drop down box at the bottom.

A lot of people are keen to use this type of quantitation for enrichment type experiments such as ChIP-Seq and would use it to normalise against an input sample. Whilst this is a common procedure, you should be somewhat cautious about applying it. When you normalise against a reference the final accuracy of your normalised value is dependent on the accuracy of both the enriched and the reference samples. In a ChIP experiment the enriched sample is generally measured very well since the reads will cluster in the measured regions, but the input sample is often poorly measured since the reads are evenly spread over the whole genome. This means that the normalised values are more likely to be influenced by technical variation in the input measures and that normalising may make your data worse. As an alternative you could consider using the input samples as a filter, to remove places where the input shows significant enrichment, and which are therefore also likely to give inaccurate measures in the enriched sample.

Manual correction quantitation

For some specialised applications it may be that the information you need to correctly normalise your data isn’t present in the mapped sequences themselves. Most sequencing applications are by their nature relative measures, telling you what proportion of reads fall into a particular gene, but not how this relates to an absolute measure in your original sample. To take a simple example, two RNA-Seq datasets taken from samples where the distribution of transcripts was identical, but the absolute level of transcription was different, would look exactly the same in the mapped data. Only by incorporating some external measure could you account for this type of difference.



To allow you to apply these kinds of correction SeqMonk has a manual correction option. This lets you enter a manual correction value for each of your datasets and choose how this is applied to the set of quantitated values calculated by SeqMonk. By using this option, you can incorporate absolute external measures with the relative measures the program itself can produce.

Working with larger data sets

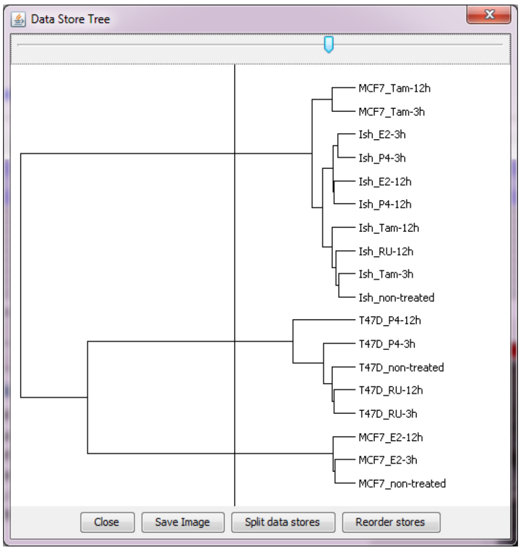
In this section we'll go though some of the functionality in SeqMonk which can make it easier to work with larger number of samples.

Data Store Clustering

When you have large collections of samples one of the earliest analysis to do after quantitation is to cluster your samples. The reason for doing this is to see if the biological groups you expect in your data emerge naturally from the results, or if there is any other structure to your data which you'd want to understand before going any further with your analysis.

SeqMonk has a few different tools to use for looking at the overall relatedness of data sets.

The Data Store Tree

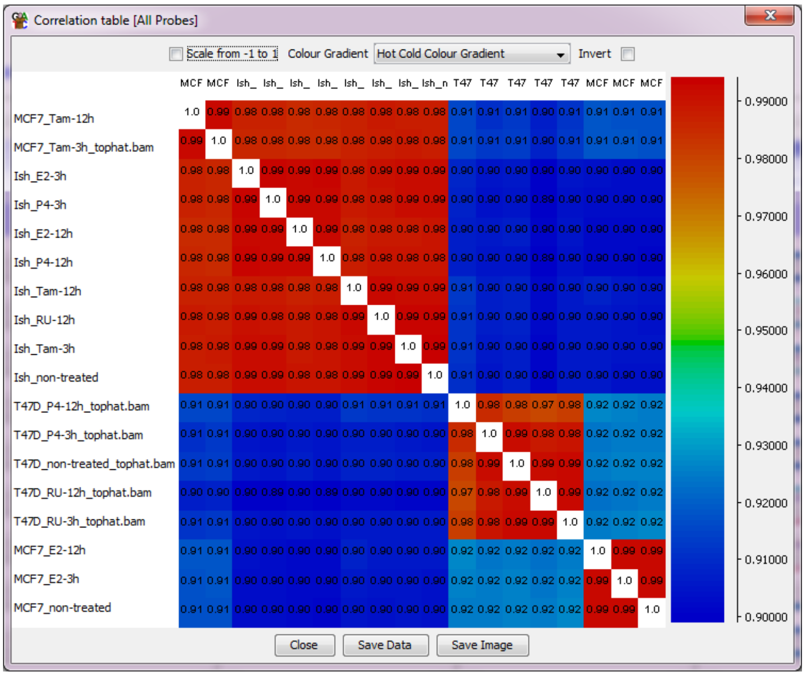


One of the simplest ways to look at the relatedness of different data stores is to do an all-vs-all correlation of the samples. You can actually generate a correlation matrix in SeqMonk from Plots > Data Store Similarity > Correlation matrix, but this is difficult to interpret. An easier way to view this data is to plot it out as a tree, which you can do with Plots > Data Store Similarity > Data Store Tree. In this view the correlation between samples is translated into branch lengths, so that samples which cluster together more tightly appear as sub-groups on the tree.

From this tree view you can choose to re-order the tracks in the chromosome view to match the clustering in the tree so that data stores which are most similar are put next to each other. You can also use the slider at the top to partition the tree into replicate sets so that similarly behaving samples can be grouped together without any prior knowledge.

The Correlation Matrix

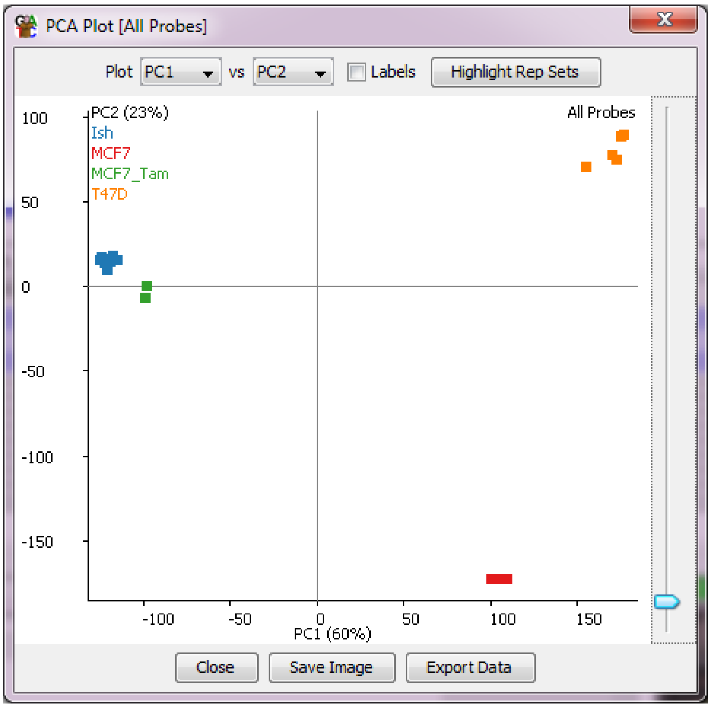
To get a more detailed view of the data which went into the data store tree you can plot out a matrix of correlation values between all pairs of samples. This will give you the quantitative values which were used by the data store tree.



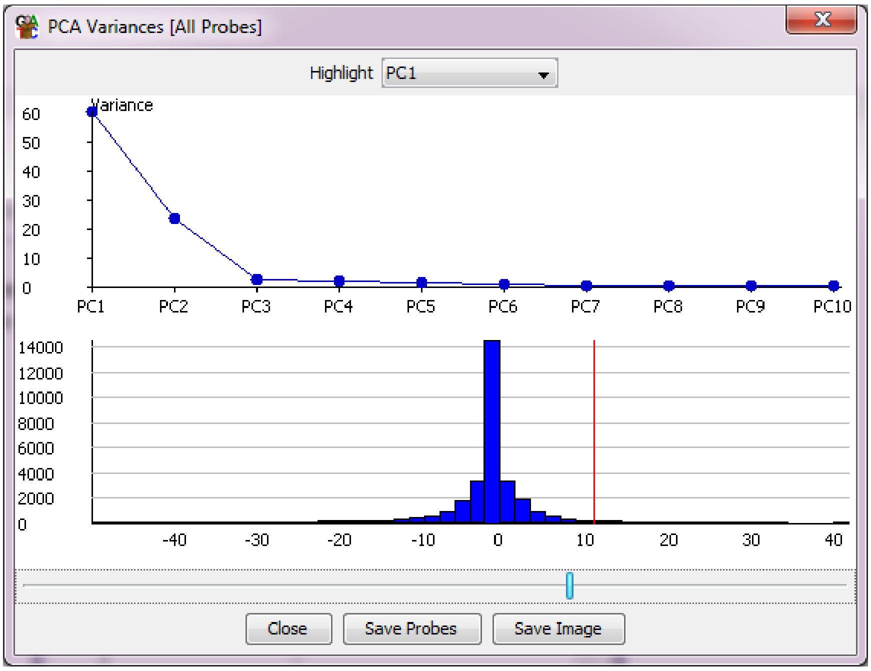
	MCF	MCF	Ish_	Ish_	Ish_	Ish_	Ish_	Ish_	Ish_	Ish_n	T47	T47	T47	T47	T47	MCF	MCF	MCF
MCF7_Tam-12h	1.0	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.91	0.91	0.91	0.90	0.91	0.91	0.91	0.91
MCF7_Tam-3h_tophat.bam	0.99	1.0	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.91	0.91	0.91	0.90	0.91	0.91	0.91	0.91
Ish_E2-3h	0.98	0.98	1.0	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
Ish_P4-3h	0.98	0.98	0.99	1.0	0.99	0.99	0.98	0.98	0.99	0.99	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90
Ish_E2-12h	0.98	0.98	0.99	0.99	1.0	0.99	0.98	0.98	0.98	0.98	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
Ish_P4-12h	0.98	0.98	0.99	0.99	0.99	1.0	0.98	0.98	0.98	0.98	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90
Ish_Tam-12h	0.98	0.98	0.98	0.98	0.98	0.98	1.0	0.99	0.99	0.99	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90
Ish_RU-12h	0.98	0.98	0.99	0.98	0.98	0.98	0.99	1.0	0.99	0.99	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90
Ish_Tam-3h	0.98	0.98	0.99	0.99	0.98	0.98	0.99	0.99	1.0	0.99	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90
Ish_non-treated	0.98	0.98	0.99	0.99	0.98	0.98	0.99	0.99	0.99	1.0	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90
T47D_P4-12h_tophat.bam	0.91	0.91	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.91	1.0	0.98	0.98	0.97	0.98	0.92	0.92	0.92
T47D_P4-3h_tophat.bam	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.98	1.0	0.99	0.98	0.98	0.92	0.92	0.92
T47D_non-treated_tophat.bam	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.98	0.99	1.0	0.99	0.99	0.92	0.92	0.92
T47D_RU-12h_tophat.bam	0.90	0.90	0.90	0.89	0.90	0.89	0.90	0.90	0.90	0.90	0.97	0.98	0.99	1.0	0.99	0.92	0.92	0.92
T47D_RU-3h_tophat.bam	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.98	0.98	0.99	0.99	1.0	0.92	0.92	0.92
MCF7_E2-12h	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.92	0.92	0.92	0.92	0.92	1.0	0.99	0.99
MCF7_E2-3h	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.92	0.92	0.92	0.92	0.92	0.99	1.0	0.99
MCF7_non-treated	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.92	0.92	0.92	0.92	0.92	0.99	0.99	1.0

The PCA Plot

Another way of looking at the relatedness of multiple samples is to use a principal components analysis (PCA). A limitation of correlation based analyses is that you can only come up with a single overall value of the relatedness of two samples, but in some cases different subsets of probes will give different structure to the relationship between samples. A PCA analysis generates a set of principal components, which are simply a set of transformations applied to all probes. Each principal component can be used to separate the samples, and different components will explain different amounts of the overall variability so that where multiple factors are relevant a PCA plot will help you to dissect these. The results of a PCA are traditionally displayed on a 2D scatterplot.

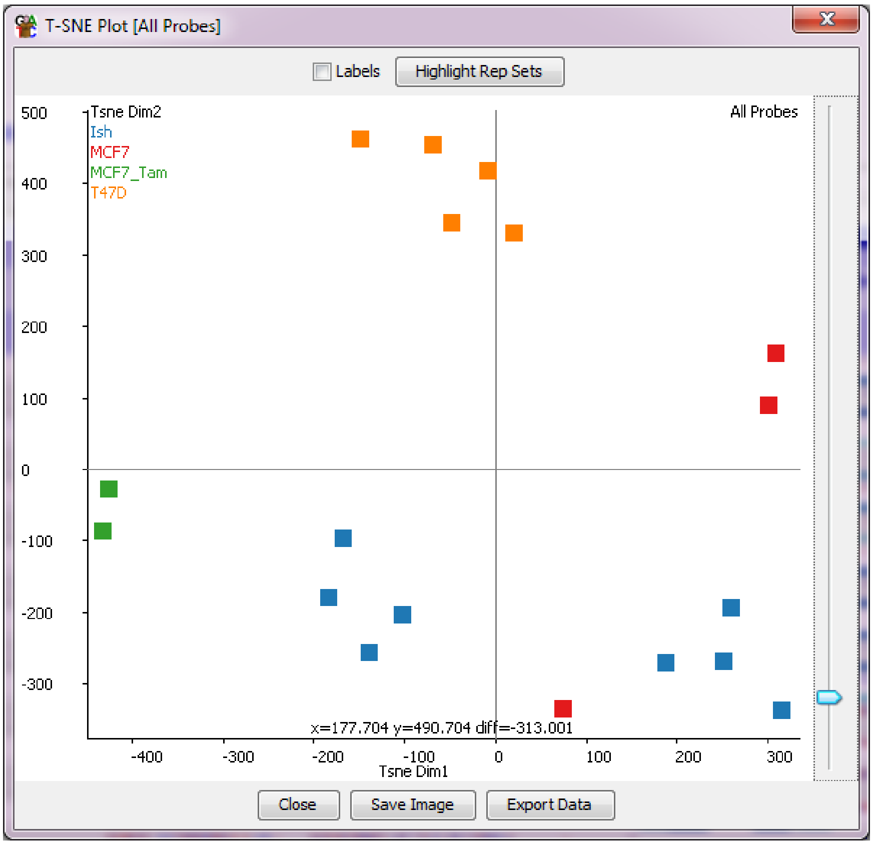


In addition to the scatterplot a second window allows you to look at the set of rotations (transformations) for a given principal component. From here you can select probe with unusually high or low rotations, which are likely to be the ones which are most influental in this component.



The Tsne Plot

Tsne plots fulfil a very similar role to PCA plots, but with a few differences. They use a more advanced machine learning algorithm which can often cope better with the complex relationships which exist in larger datasets. The separation you see is often therefore more visually pleasing. In a Tsne plot you only ever get 2 dimensions of separation so the option to show the different components is not present. Also you can't extract from a Tsne plot which probes were the main drivers of a particular separation, so it's more for visual inspection than more rigorous analysis.



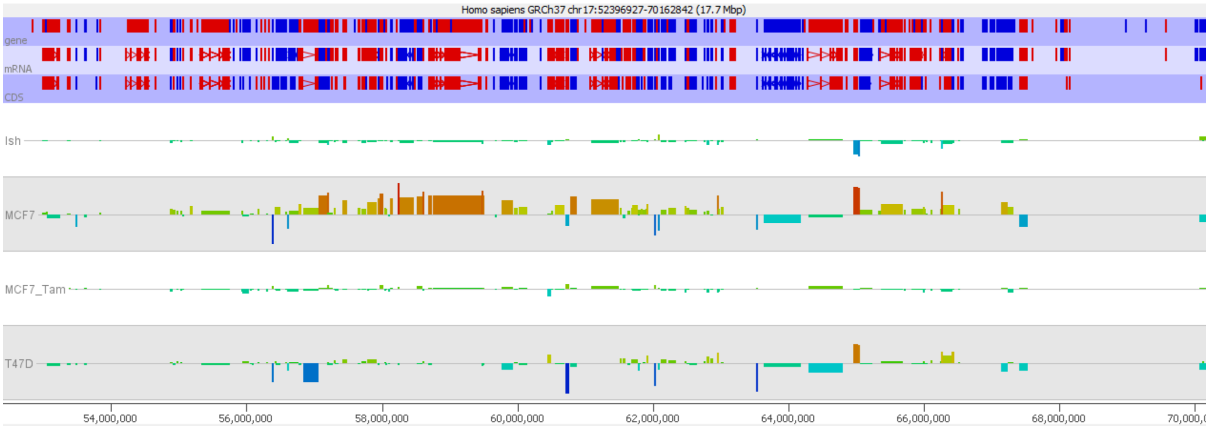
x=177.704 y=490.704 diff=-313.001

View options for larger studies

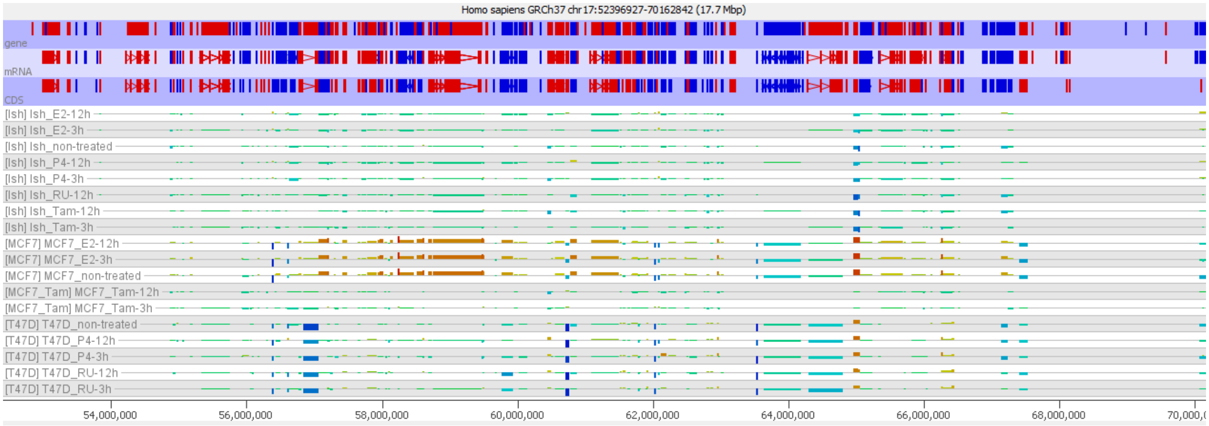
When a larger number of samples is imported you will find that some of the standard views in the chromosome view because cluttered and uninformative. To work round this there are some features which are specifically designed to make working with large sample numbers easier.

Different ways to represent Replicate Sets

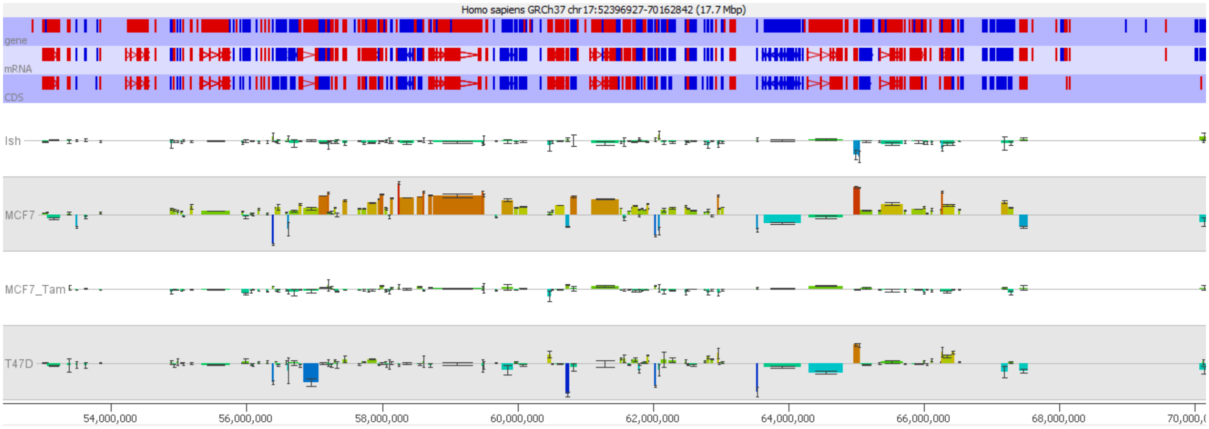
When dealing with replicate sets containing large numbers of samples there are some options for how these are shown in the chromosome view. By default, replicate sets are displayed by showing the merged set of raw reads, and the mean values for any quantitation.



Under View > Data Track Display there are a couple of options to say how to display replicate sets. The “Replicate Set Display” setting allows you to expand a replicate set track into its sub-components. These will show up as individual tracks in both the chromosome view and any displays which take their data from the visible data stores, but it will also allow the set of tracks to be manipulated as a single object in the Data Track Display options.

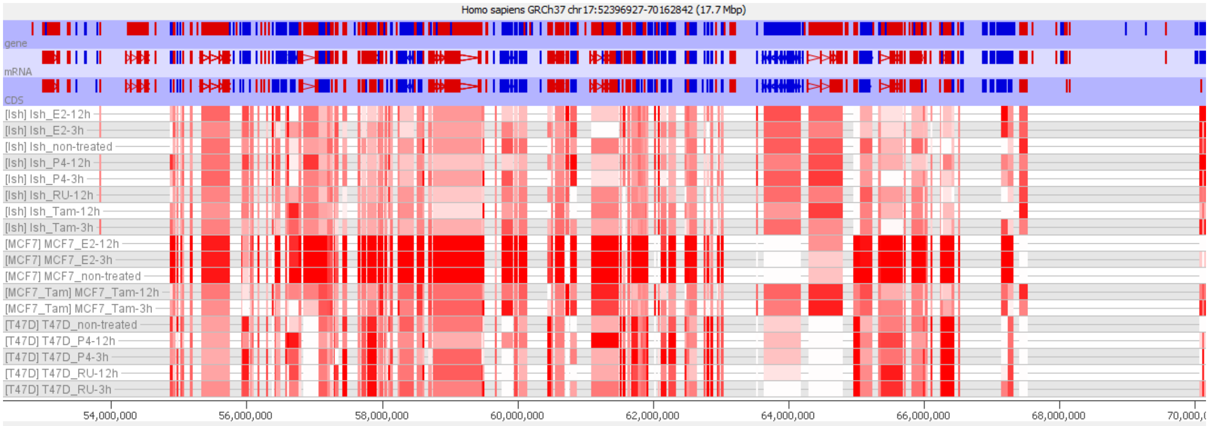


If you choose to show replicate sets as a single track, then you can use the "Replicate set Variability" option to put error bars onto the measures shown. You can choose from StDev, SEM, individual points or high/low values.



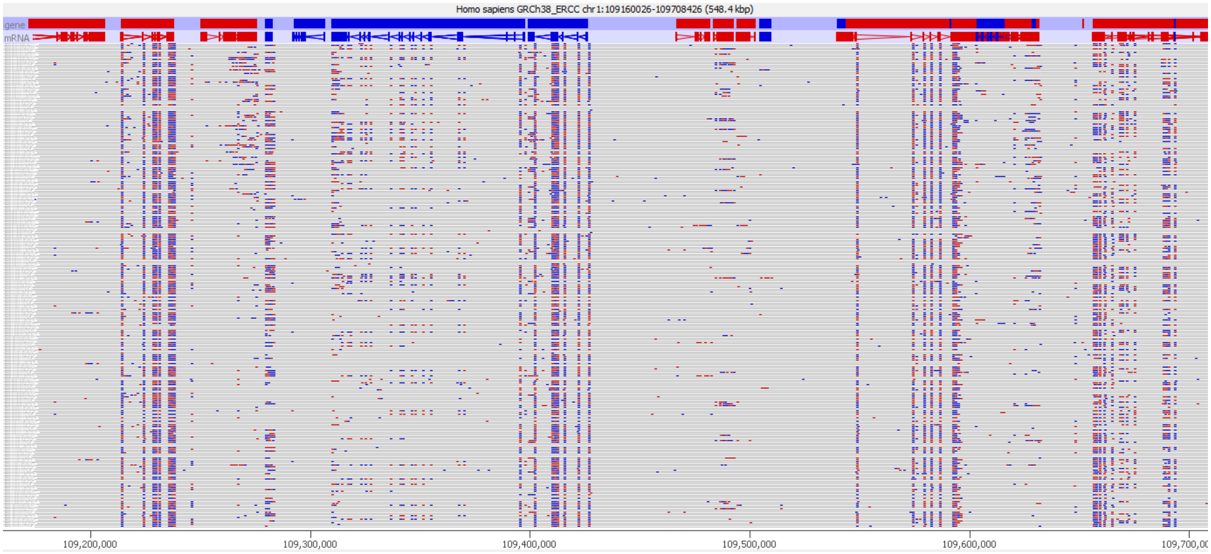
Quantitated data plotting for large sample numbers

The display preferences has a number of options for how quantitated values are shown in the chromosome view. Most of these rely on using the height up a track as the way to show the quantitated value, but this does not scale well for large sample numbers. In these cases, you can switch to using a colour based scaling, where each probe appears simply as a block of colour, but the colour of the block is scaled to the current quantitation. This type of display effectively turns the chromosome view into a large heatmap and can effectively show large numbers of samples at once.

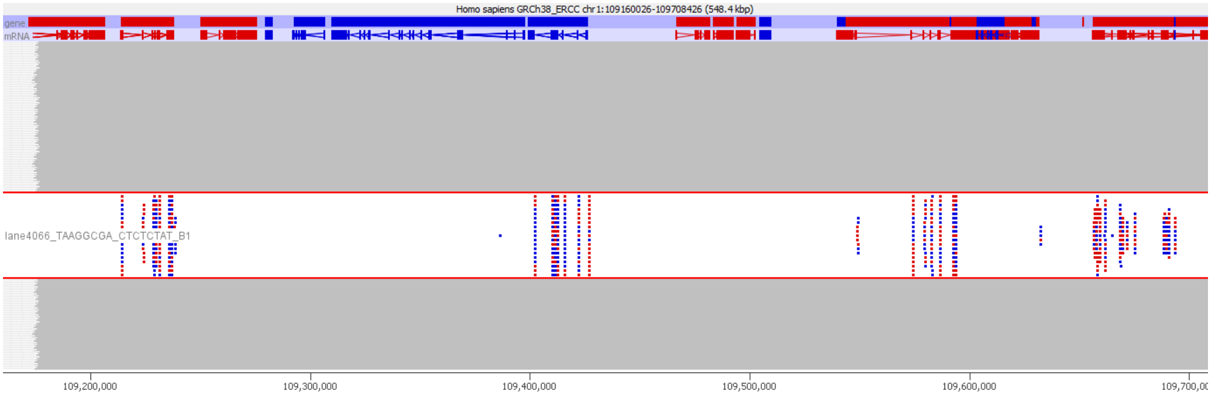


Chromosome view adjustments for large sample numbers

Once you have a large number of visible data stores in the chromosome view you lose the ability to see any detail in the raw data view. The amount of vertical space per sample is so small you have no ability to assess what's going on.



However, in these circumstances SeqMonk still offers a solution. If you select a data store from the data view window, then if the vertical space offered to that store is too small then it will re-weight the tracks to make this store taller at the expense of all of the other tracks so you can still examine the raw data.


Homo sapiens GRCh38_ERCC chr1:109160026-109708426 (548.4 kbp)
gene
mRNA
lane4066_TAAGGCGA_CTCTCTAT_B1
109,200,000
109,300,000
109,400,000
109,500,000
109,600,000
109,700,00

Statistical Filters

SeqMonk offers a large number of statistical filters. These are arranged under the Filtering > Filter by Statistical Test menu. To try to make it easier to select an appropriate test the filters are divide based on the type of analysis you're performing and the number of replicates you have available.

Test	Data Type	Replicates Required	Notes
Replicate Set Stats	Continuous	Yes	Standard t-test or ANOVA. Assumes normalised normally distributed data.
LIMMA	Continuous	Yes	Variance stabilised ANOVA - more powerful for smaller experiments. Assumes normalised normally distributed data.
Intensity Difference	Continuous	No	Statistical magnitude of effect test. Find unusually highly changing values.
Windowed Replicate	Continuous	No	Look for consistently changing, physically adjacent probes
DESeq2	Counts	Yes	Powerful count based comparison. Requires raw counts for quantitation.
EdgeR	Counts	Yes	Powerful count based comparison. Requires raw counts for quantitation.
Proportion of Library	Counts	No	Chi-Square test on proportion of reads falling into a single probe compared to the library as a whole
Logistic Regression (for/rev)	Proportions	Yes	Replicated stats based on the proportion of for/rev reads in a probe. Normally used for bisulphite analysis.
EdgeR (for/rev)	Proportions	Yes	Replicated stats based on the proportion of for/rev reads in a probe. Normally used for bisulphite analysis.
Logistic Regression (splicing)	Proportions	Yes	Replicated stats based on paired splicing counts for common splice donors/acceptors. Used to detect differential splice junction usage
Chi-Square (for/rev)	Proportions	No	Unreplicated comparison of for/rev ratios in a probe. Normally used for bisulphite analysis.
Chi-Square (multiple stores)	Proportions	No	Comparison of count ratios between paired stores. Normally used for comparing allelically separated data.
Chi-Square (front/back)	Proportions	No	Comparison of count ratios in the front / back half of a probe. Normally used to assess termination ratios in RNA-Seq data.
Binomial (for/rev)	Proportions	No	More complex comparison of for/rev ratios which works in the presence of a large global shift. Normally used for bisulphite analysis.
Variance Intensity Difference	Variance	Yes	Statistical test used to find usually high/low variance values within a replicate set.

Gene Set Enrichment	Subgroup	No	Uses a set of pre-defined lists to find subsets with unusual distributions of values. Uses t-test or KS test internally.
Monte Carlo Simulation	Subgroup	No	Does a random simulation to find whether an existing gene list has unusually biased quantitated values compared to randomly selected genes.
Box Whisker	Outlier Detection	No	A simple test to look for outlier values from data which is roughly normally distributed.

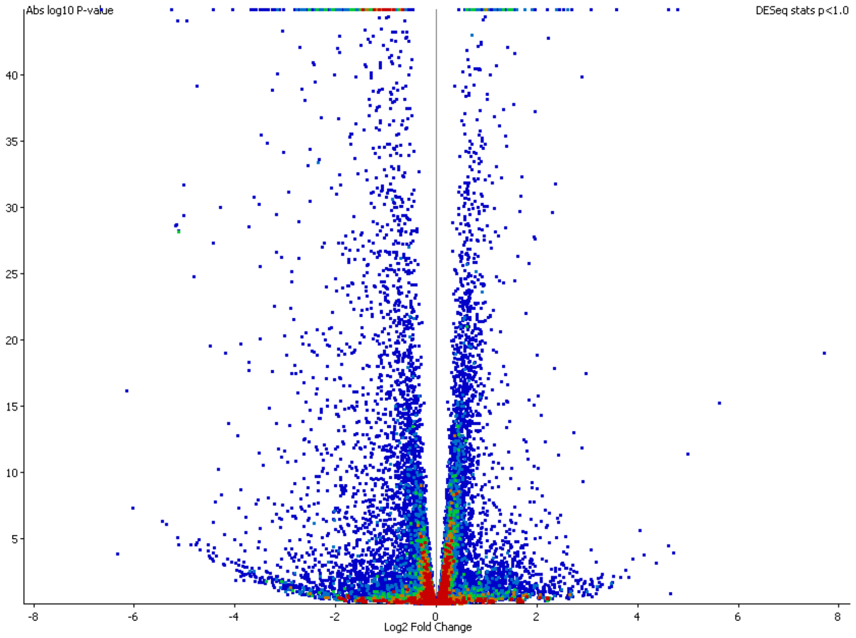
Running statistical filters

In general statistical filters operate the same as every other filter. You select the lists you want to test, run the filter and then a sub-probe list is created containing the hits.

All statistical filters will give you the option to return only the probes which were statistically significant (normally after multiple testing correction) in the sub-list. You can however also choose to run the test in such a way that all possible p-values are returned. To do this simply turn off multiple testing correction, and set the p-value cutoff to be 1, and all possible results should be returned.

Drawing volcano plots for statistical results

One nice way to look at the hits which were selected from a statistical test is to use a volcano plot. This plots out the absolute log10 p-value (so larger is more significant) against the difference value for the values being tested. You can draw a volcano plot by selecting any probe lists created by a statistical filter which records both p-value and magnitude of difference (most do, but not all of them), and then selecting Plots > Volcano Plot.



As with the other scatterplots you can mouse over a point to see what it is and double click on it to go to that point in the chromosome view.

Filtering on probe list annotation

Another related function which can be useful for statistical filtering is the ability to filter on any of the annotated values put onto a probe list. This would allow you to filter on the reported fold change from a filter as well as the p-value for example.

The annotation value filter is another one of the filters listed under Filter by Values.

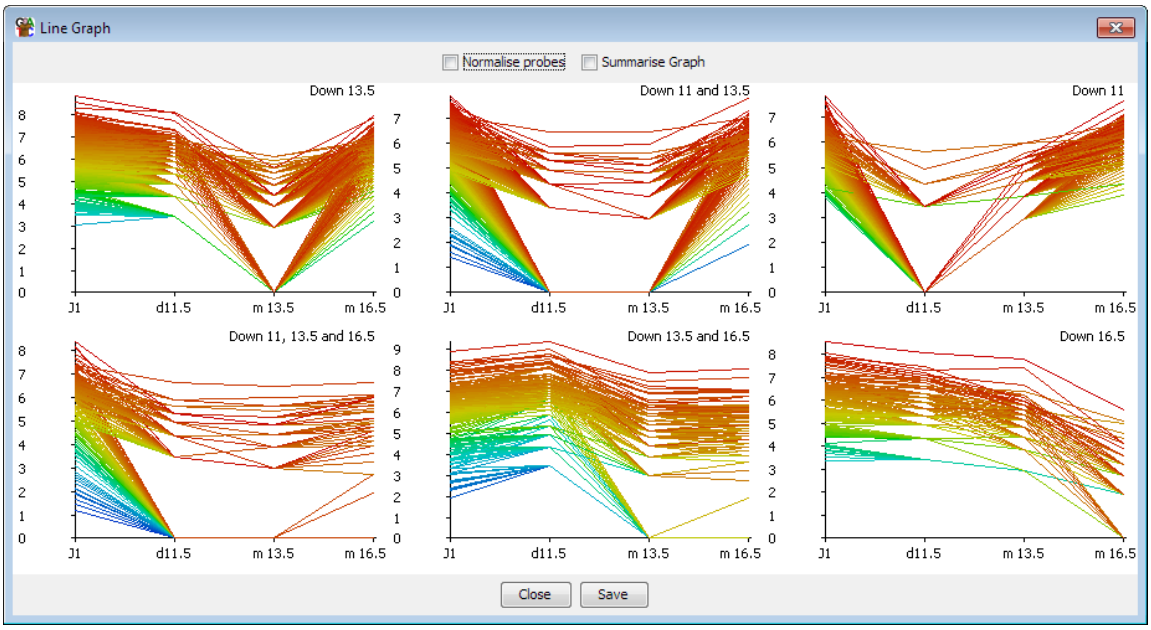
Clustering and Grouping Probes

When you have used the filters in SeqMonk to generate a subset of probes of interest you can often find that interpreting the biology behind the set of hits can still be difficult. This is especially true where you have several experimental conditions where different sub-groups of probes will behave differently across the set of conditions.

To try to make the picture clearer you can use clustering of your probes to find subsets of probes which behave similarly. This can give you several clean signals to interpret instead of one noisy one.

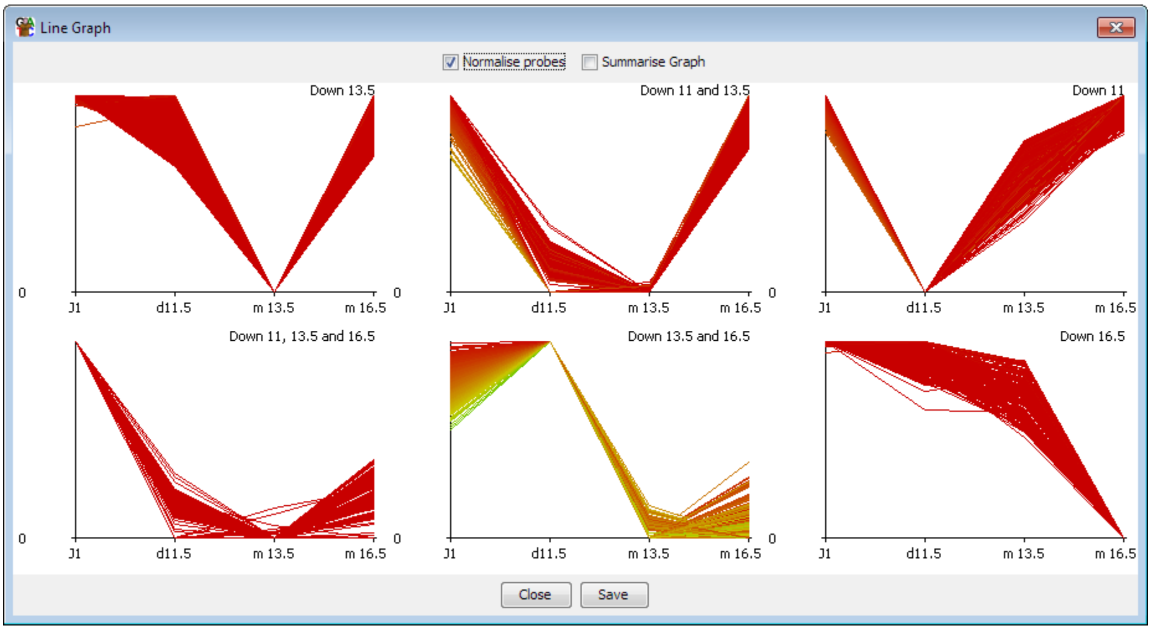
The line graph plot

Before we get into doing the actual clustering it's useful to know how the results can be displayed. SeqMonk provides the line graph plot which allows you to look at the changing quantitation of thousands of probes across multiple datasets, and which can also show several groups (clusters) at once.

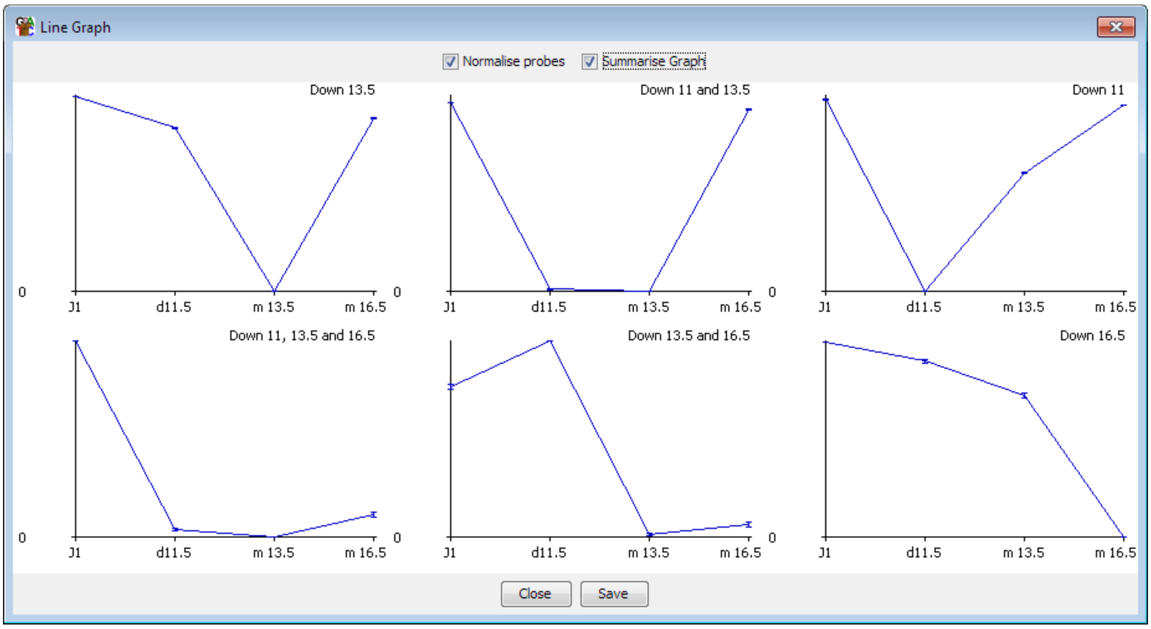


In a clustered set of probes, the shape of the lines over the various data stores should be somewhat similar. However, with large numbers of probes you may still find the plots descending into chaos. There are therefore a couple of additional options which can help to clean things up.

Clustering is based on observing similar patterns of change between conditions, however when the same pattern of change is observed at different absolute levels of quantitation you end up with a line graph containing multiple parallel lines. With large numbers of probes these parallel lines merge into a block, such that it is hard to discern the overall pattern observed in the group. One way to clean up the view is therefore to perform a per-probe normalisation. For each probe this normalisation subtracts the median value across all data stores from each individual point. In effect it adjusts the y axis for each probe so that it centres on a value of 0. What you are then comparing is the variation in the quantitated values between conditions, regardless of the absolute level of quantitation. The plot should therefore be more consistent between the different probes and the overall pattern should emerge more clearly. This per-probe normalisation can be performed in real time in the line graph plot, but should you wish to you can also make the same change to your raw data by using the per-probe normalisation quantitation method.

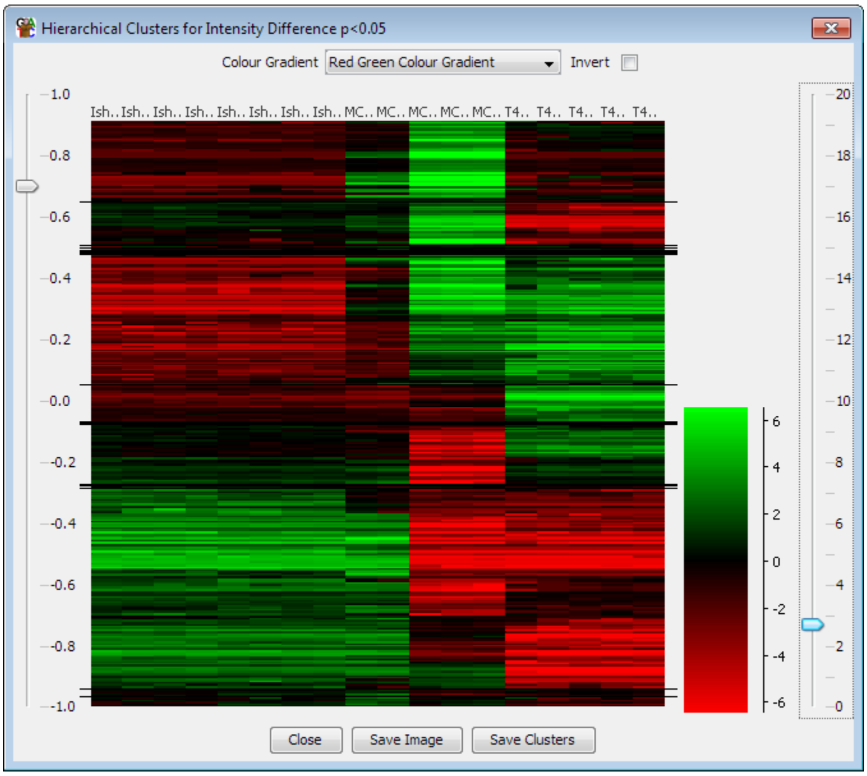


If the plot is still somewhat confused even after per-probe normalisation then a second additional option to clean up the plot is to just show the average value for each dataset, rather than showing each individual point. In this mode a single line is shown, but for each point a set of error bars indicate the standard deviation of the values in that data store. This should provide a good view of the overall pattern across several samples even where there is a high level of variability.



Hierarchical clustering

If you have a relatively small number of probes to cluster (fewer than 5000 would be a sensible limit) the best place to start looking at clustering is the hierarchical clustering tool. This is a visual clustering where a set of probes is rearranged so that probes with a similar quantitation profile across your set of conditions are placed closely together.


Hierarchical Clusters for Intensity Difference p<0.05
Colour Gradient
Red Green Colour Gradient
Invert □
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Ish.. Ish.. Ish.. Ish.. Ish.. Ish.. Ish.. Ish.. MC.. MC.. MC.. MC.. MC.. T4.. T4.. T4.. T4.. T4..
6
4
2
0
-2
-4
-6
20
18
16
14
12
10
8
6
4
2
0
Close
Save Image
Save Clusters

As you can see there are obvious groupings of probes within the plot. If you want to be able to define these groups, then you can use the slider on the left to divide the plot into groups which show a level of correlation above a threshold you specify. You can then save these groups as probe lists in your project. You can zoom into the y-axis of the plot by dragging box within the plot area. When you export lists of probes only groups which are currently visible in the plot are exported.

Automated Correlation Clustering

Another place to start working with clusters if you have a large number of probes is to perform an automated correlation clustering. Automated clustering requires at least 3 different data stores to work with and is based on grouping together probes whose quantitation patterns across those stores shows a high Pearson's correlation coefficient. It is implemented as a probe list filter and can be found under Filtering > Filter by Correlation > Filter by Correlation Cluster.

When setting up the clustering you should start with a set of probes you already know to be changing between your conditions. Including unchanging probes will start to cluster on the noise in your system and your results will be poor and take ages to calculate.

The process for automated clustering is that the first probe analysed starts a cluster. Every probe after that is correlated to the existing clusters. If it can be placed in an existing cluster, then it is. If not, then it gets to start a new cluster. Finally, when all probes have been clustered you can apply a filter to remove clusters with very low numbers of probes in them.

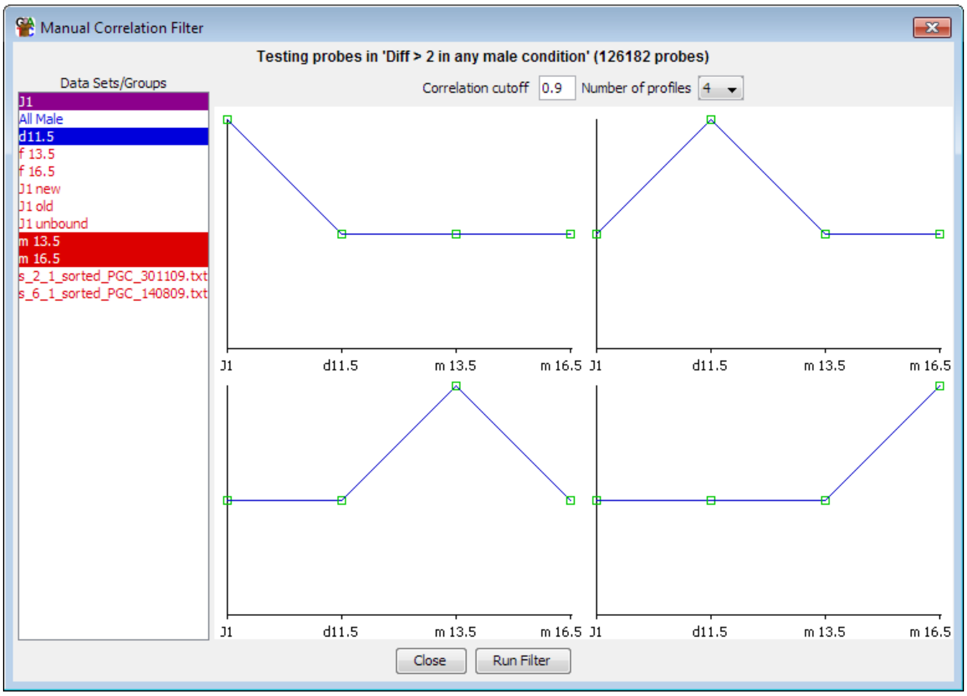
The stringency of clustering is based on the degree of correlation you want to see within each cluster. Setting a high correlation cut-off (e.g. 0.9) will produce lots of small clusters with very similar profiles. Using a low cut-off (e.g. 0.7) will produce fewer, larger clusters but with more variability in the profiles contained within an individual cluster.

Automated clustering can be a useful step in analysing your data, but the results it produces will depend on the order in which probes are passed to it. The results can therefore be somewhat noisy – you may sometimes see two or more clusters produced which have very similar overall profiles. The clusters produced by this automated clustering should therefore not be taken as a final answer but will give you an idea of the breakdown of your probes, and once you know this you can use the manual clustering options to get a cleaner set of results for the interesting clusters you find.

Manual clustering

Rather than performing an unbiased automated clustering, the other option you have is to do a manual clustering. In a manual clustering you define a starting set of profiles you want to find. Every probe is then correlated against this set of profiles – any probe which fails to correlate to any of the profiles is rejected and those that do correlate are placed in a group with the profile to which they correlated most tightly.

Unlike automated clustering which can give noisy and inconsistent results, the results of manual clustering should be cleaner and reproducible, but you will only find the profiles you specified to start with.


Manual Correlation Filter

Testing probes in 'Diff > 2 in any male condition' (126182 probes)

Correlation cutoff 0.9 Number of profiles 4

Profiles for manual clustering can be specified in one of two ways. You can choose to create a profile manually, in which case you are presented with a list of your currently visible stores and you can manually drag the control points on the profile shown to create the profile you want to see. This gives you complete control over the shape of the profile you want to find. Alternatively, you can select an existing probe list and the clustering will create an averaged profile from the data in this list and then cluster against this profile. The lists you select can contain any number of probes, so you could, for example, make a list containing a single probe if you wanted to find other probes which were similar to a probe of interest.

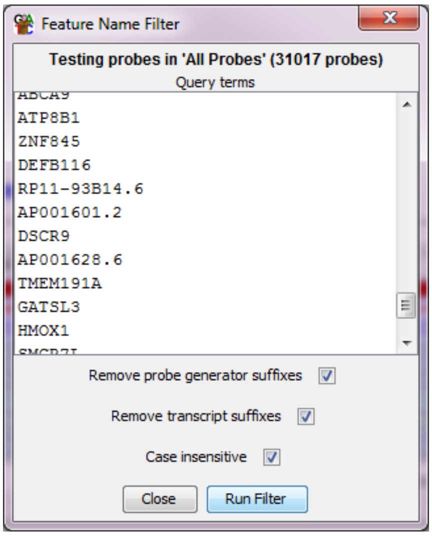
Linking SeqMonk to external programs

Whilst SeqMonk offers a wide range of quantitation, visualisation and analysis tools there will always be cases where a specific analysis isn't available within the program. In these cases, it is useful to know how to export and import information to an external source. If you do find what you think might be a common additional requirement in SeqMonk then do please let us know about it so we can see if we can include it in a future release.

Exporting and importing probes and quantitations

We saw earlier on how you could use reports to export lists of probes with their corresponding quantitations into a text file which would be suitable to transfer into another analysis platform. There are a number of ways you can bring results back into SeqMonk.

Importing hits from an external statistical test



Feature Name Filter

Testing probes in 'All Probes' (31017 probes)

Query terms

ABCA9
ATP8B1
ZNF845
DEFB116
RP11-93B14.6
AP001601.2
DSCR9
AP001628.6
TMEM191A
GATSL3
HMOX1

Remove probe generator suffixes ☑

Remove transcript suffixes ☑

Case insensitive ☑

If you have generated quantitations in SeqMonk and you then produce a subset of these as hits from an external tool, you can import these back into SeqMonk for further investigation. There are a few ways to do this.

If your probes have unique names (and they often will) then you can just take the set of names for the hits and use the "Filter by Probe Names" option. You can paste in the names of the hits and the corresponding probes will be put into a new Probe list. The same strategy works if you have a set of gene names as hits and gene based probes, where the Probe names will also be the gene names.

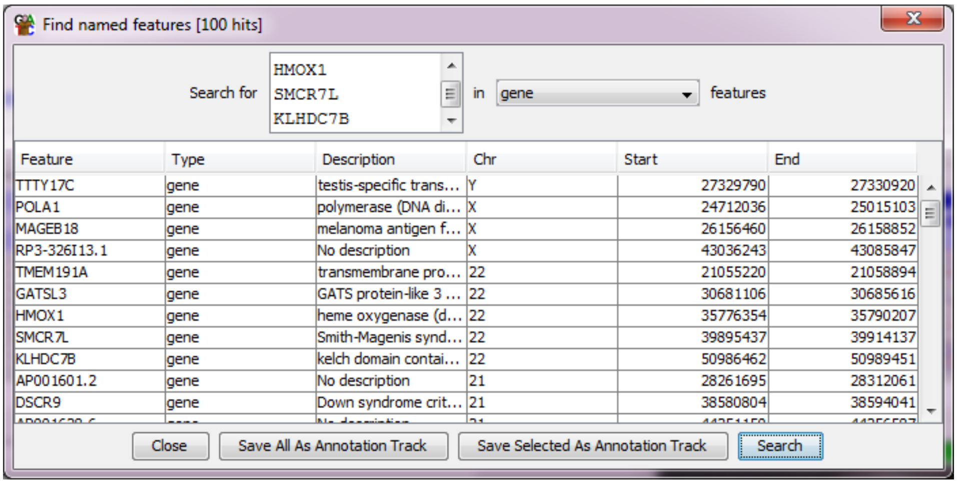
If you can't use the names of the probes then you can import the positions of the hits as a new annotation track (File > Import Annotation > Text Generic), and then use the Features filter with the "Exactly Matching" relationship option to select the corresponding probes.

Relating probes to an external position list

If you have externally generated a set of interesting positions (results of a peak caller for example), you can bring these into SeqMonk as an annotation track (File > Import Annotation). If you want to relate these to an existing set of probes you can use the Features filter to find probes which are close to or overlapping with these external positions without requiring an exact match. If you want to define probes around these positions, then you can use the Feature Probe Generator to do this.

Relating probes to an external feature list

If you have a set of feature names (gene names for example), but you don't have the positions, or if you have the names of multi-locus positions (mRNA for example), then you can bring these into SeqMonk by sub setting an existing feature track. To do this you can use Edit > Find Named Features and then paste in your feature list, selecting the appropriate SeqMonk feature track to filter against. Once you have the list of hits you can use buttons at the bottom of the dialog to turn the hits into a new annotation track which you can then use with Feature filters or Feature probe generators.


Find named features [100 hits]
Close
Save All As Annotation Track
Save Selected As Annotation Track
Search

Search for: HMOX1, SMCR7L, KLHDC7B in gene features

Feature	Type	Description	Chr	Start	End
TTTY17C	gene	testis-specific trans...	Y	27329790	27330920
POLA1	gene	polymerase (DNA di...	X	24712036	25015103
MAGEB18	gene	melanoma antigen f...	X	26156460	26158852
RP3-326I13.1	gene	No description	X	43036243	43085847
TMEM191A	gene	transmembrane pro...	22	21055220	21058894
GATSL3	gene	GATS protein-like 3 ...	22	30681106	30685616
HMOX1	gene	heme oxygenase (d...	22	35776354	35790207
SMCR7L	gene	Smith-Magenis synd...	22	39895437	39914137
KLHDC7B	gene	kelch domain contai...	22	50986462	50989451
AP001601.2	gene	No description	21	28261695	28312061
DSCR9	gene	Down syndrome crit...	21	38580804	38594041

Transferring locations between programs

One of the big advantages of high throughput data generation projects such as ENCODE and 1000 genomes is that there are now very large data repositories we can use to see what other information is available for a region of interest in our data. These data sets are commonly exposed in online genome browsers such as Ensembl and UCSC browser.

It can be really useful to view a region of interest in one of these browsers with some of these additional tracks turned on but this requires shifting between SeqMonk and the browser environment.

Putting quantitated data into another genome browser

One way to achieve this link is to export the current quantitation from SeqMonk and upload it to the other genome browser. The mechanism to do this is to create a bedGraph file from the current quantitation. To do this you use File > Export Current View > BEDGraph. You will be asked for a file prefix, and then a separate BEDGraph file will be generated for each currently visible data store. Alongside the BEDGraph files you will also get a chromosome sizes file so that if you want you can run the UCSC bedGraphToBigWig program to generate bigWig files. BEDGraph files can be uploaded to many genome browser systems, and bigWig files can be placed on a web server and then linked to the external source by URL.

Quickly matching positions between SeqMonk and a Genome Browser

The other way to link these two environments is to quickly match the positions of the two systems you're working with. To do this SeqMonk provides a quick way to both export and import positions in a standard `chr:start-end` format used by most of these systems.

To export the current position from SeqMonk just press Ctrl+C (or Edit > Copy Current Position). This will copy the location string of the currently visible genomic position onto the clipboard. You can then paste this into the location box on the genome browser to get to the equivalent position.

To go the other way simply copy the location string in the genome browser and then press Ctrl+G in SeqMonk (or Edit > Goto Position) and paste the string into the location box and press return.