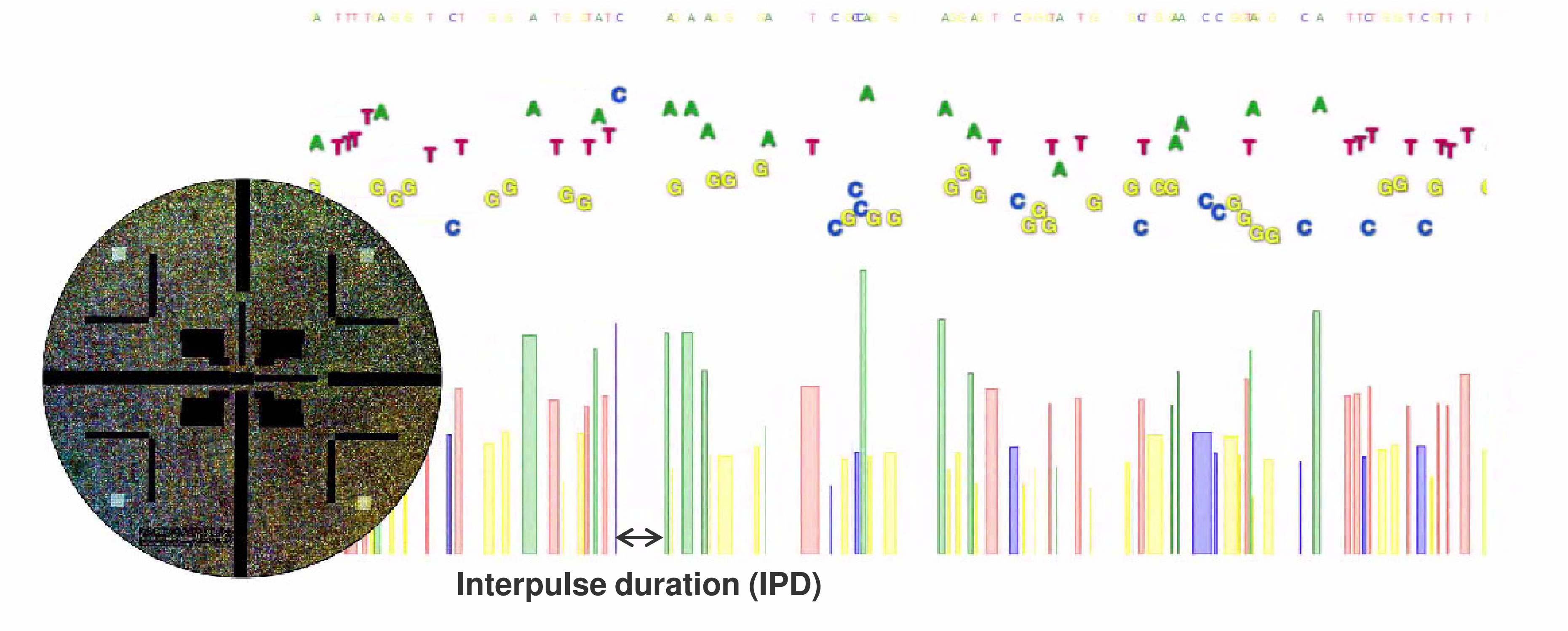
Base Modification :

From Sequencing Data to a High Confidence Motif List

Meredith Ashby

FIND MEANING IN COMPLEXITY

Kinetics in SMRT® Sequencing

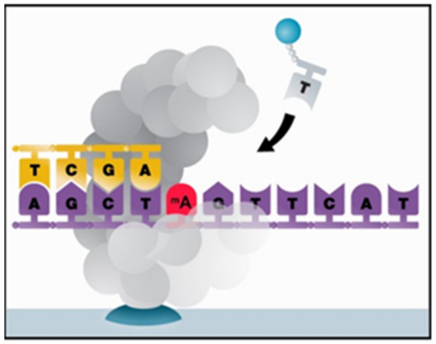


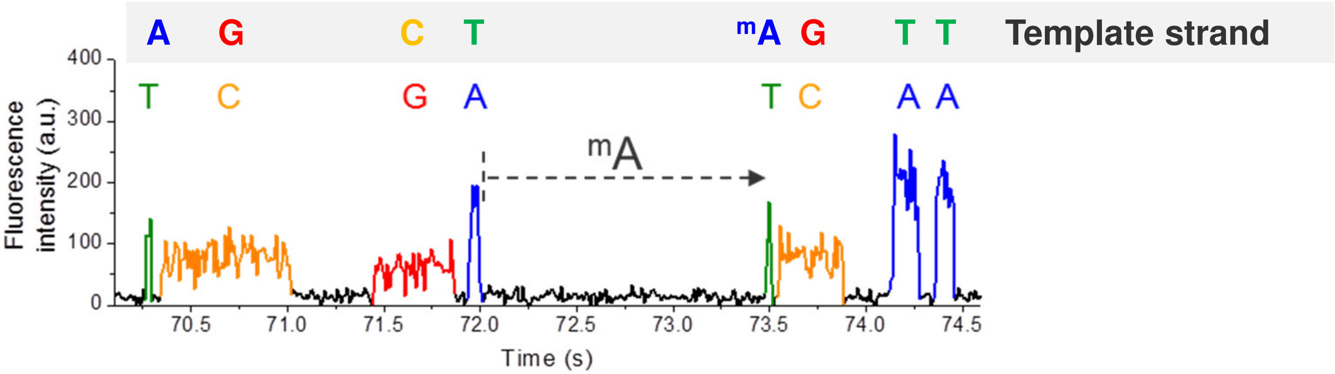
Interpulse duration (IPD)

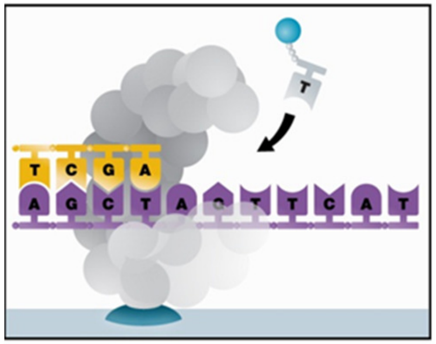
- SMRT Sequencing uses kinetic information from each nucleotide addition to call bases
- This same information can be used to distinguish modified and native bases
- We can compare results of SMRT Sequencing to an *in silico* kinetic reference for incorporation dynamics without modifications to infer the presence of bases different from A, C, G or T

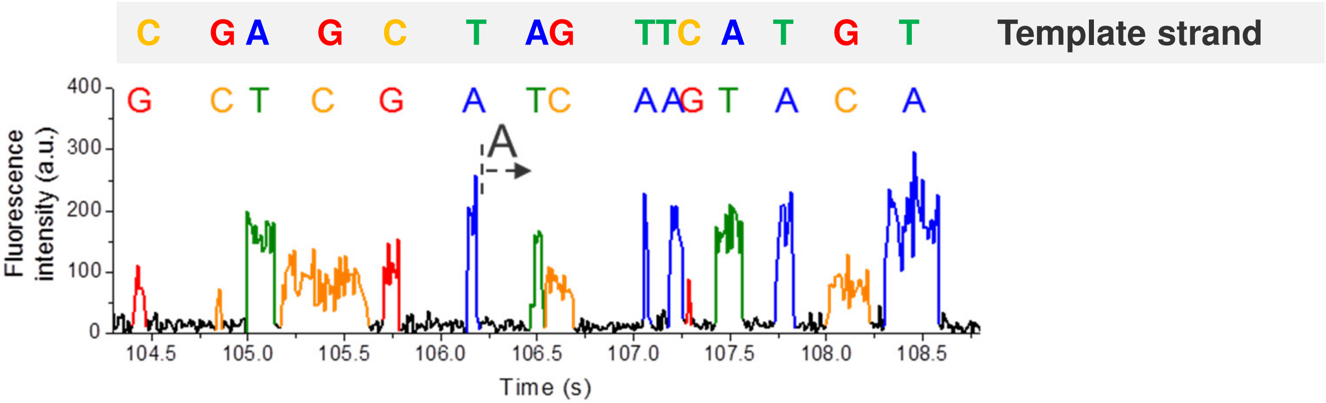
Detection of DNA Base Modifications Using Kinetics

Example: N^6-methyladenine





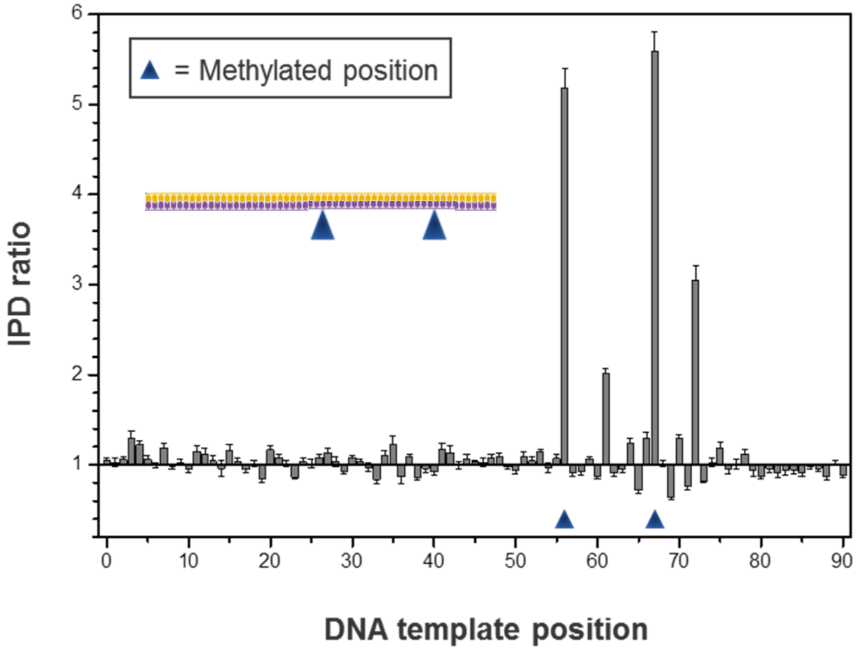


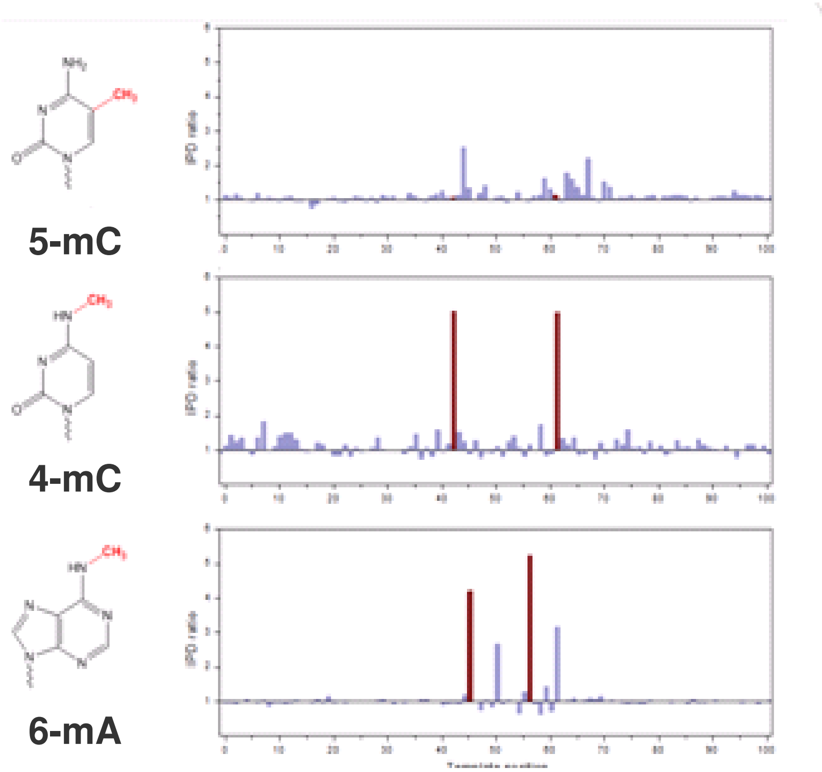


Flusberg et al. (2010) *Nature Methods* 7: 461-465

Detection of DNA Base Modifications by SMRT Sequencing

Calculation of IPD ratios across the reference gives information about base modification at every position.





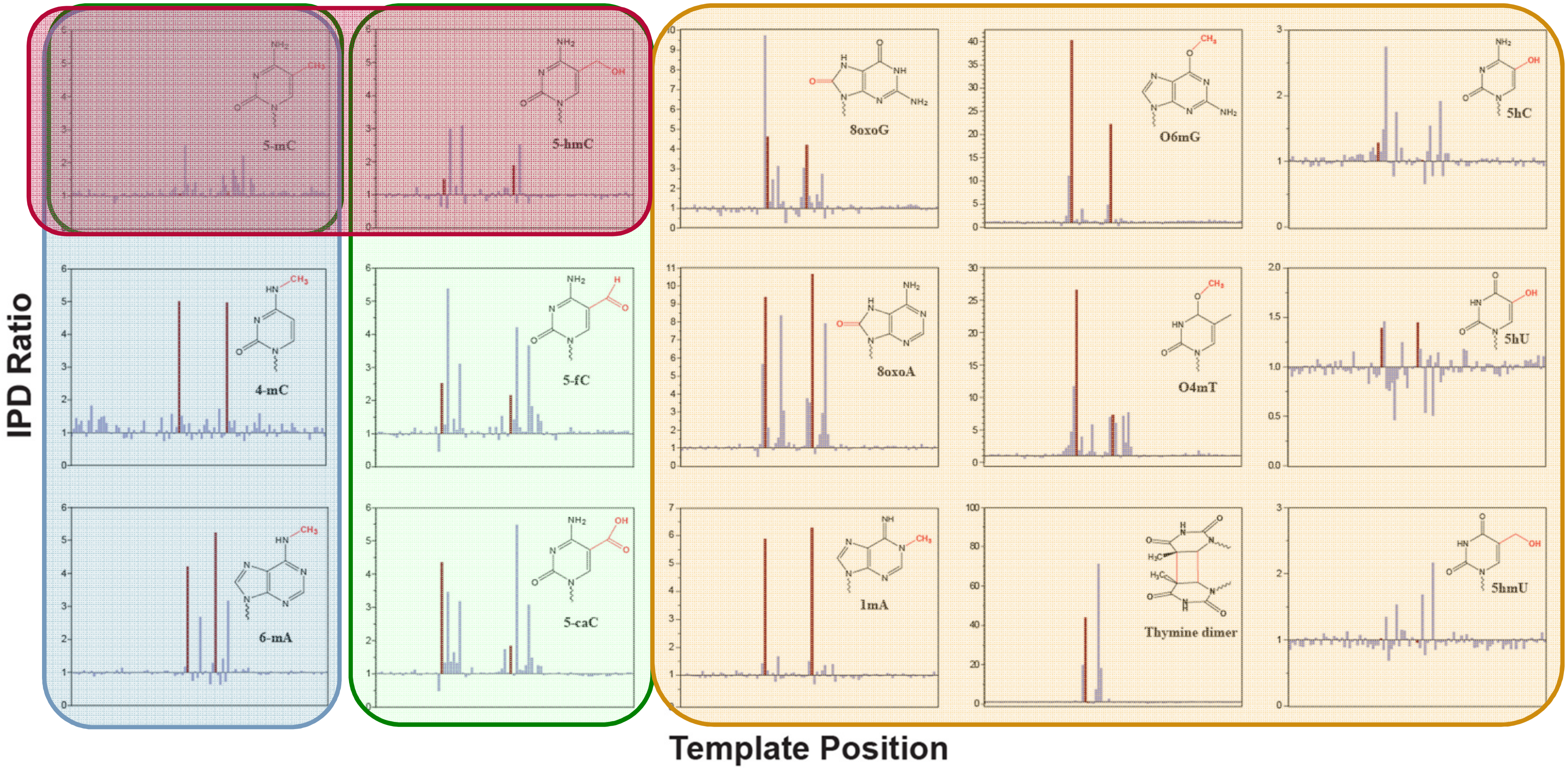
SMRT Portal v1.3.3+ can recognize and annotate multi-site modified-base signatures.

Flusberg et al. (2010) *Nature Methods* 7: 461-465

Signatures of Different DNA Base Modifications

Prokaryotic Eukaryotic DNA Damage

Detectable by other Sequencing Methods



Experimental Design For Bacterial Base Modification Projects

Experimental Design

Isolate DNA

Template Preparation

Sequencing

Analysis

Project Goals:

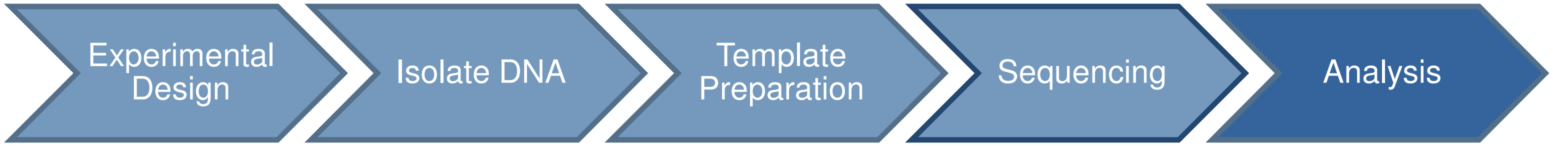
- Characterization of methylome?
  - *De novo* assembly
  - Resequencing
- Which modifications of interest?
- Identifying highly modified motifs throughout the genome vs. interrogating specific regions of the genome with high confidence?

Coverage Needs:

- Coverage needs vary based on the strength of the kinetic signal
- Kinetic signal strength varies by modification type
- Recommendation: 100x Coverage

Modification Type	Target Coverage per Strand
4mC	25x
6mA	25x
5mC	250x
TET-converted 5mC	25x

Bacterial Methylome Analysis Recommendations

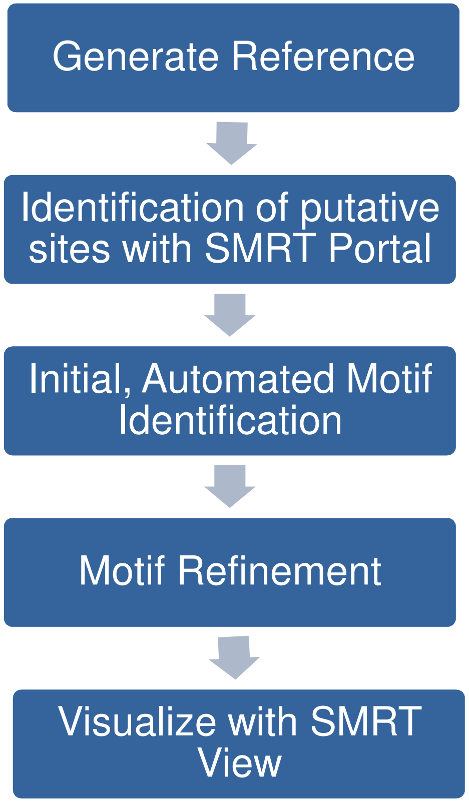


Generate Reference

- For *de novo* assembly, you will need 60-100x coverage for HGAP
- Upload Reference into SMRT Portal
- For most cases, use the *in silico* reference

Identification of putative modification sites

- SMRT Portal v2.1 or v2.2 using *RS_Modification_and_Motif_Analysis*
- Once the job is complete, evaluate whether the minimum modification QV needs to be adjusted from the default value
- If needed, refine the motif list by either:
  - Rerunning the job with the appropriate min QV setting
  - Using Motif Maker with different min QV settings
  - Using R to refine the list of 'hits' used for motif finding


Generate Reference
Identification of putative sites with SMRT Portal
Initial, Automated Motif Identification
Motif Refinement
Visualize with SMRT View

Base Modification Workflow Output

Several output files are generated by the analysis protocol

- *modifications.csv* file:
  - Comma-separated values (CSV) file with statistical analysis of each position in the reference
  - Intended to allow additional follow-up analysis for every genomic position

- *modifications.gff* file:
  - General Features Format (GFF) file
  - Used for motif analysis and modification visualization in SMRT® View
  - Includes sequence contexts for sites of putative modification
    - positions where the inter-pulse duration (IPD) is significantly different from the expected background
    - *p*-values of 0.01 or less (QV > 20)

- *motif_summary.csv* file:
  - Comma-separated vales (CSV) file with the information displayed in the Motifs report

- Files can be downloaded from the DATA section of the SMRT Portal Job Details Page

Theory vs Practice: Generating a High-confidence List of Motifs

Only include high confidence hits in the data you submit to the motif finding algorithm

- Including false positives in your data set may result in the ‘discovery’ of bizarre motifs.
- An exception is suspected 5mC motifs, which will all be low-confidence in un-Tet1 treated samples

Apply knowledge about SMRT Sequencing base modification signals and PacBio motif finding algorithm

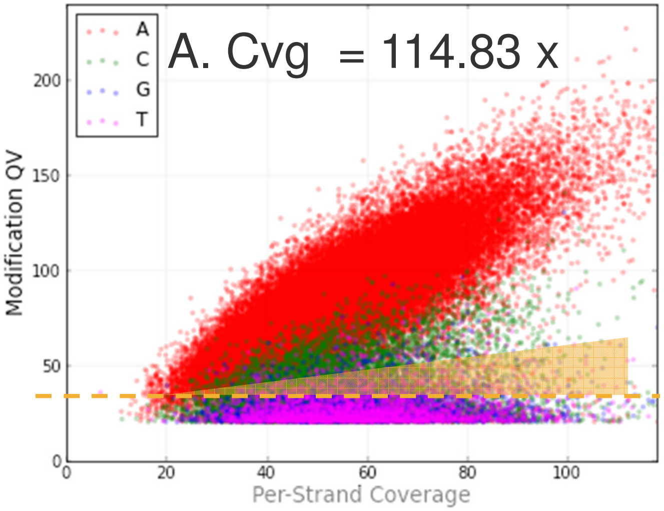
- m6A bases give significant signal 5 bases upstream of the modified A
- 5mC bases give the strongest signal 2 bases upstream of the modified C

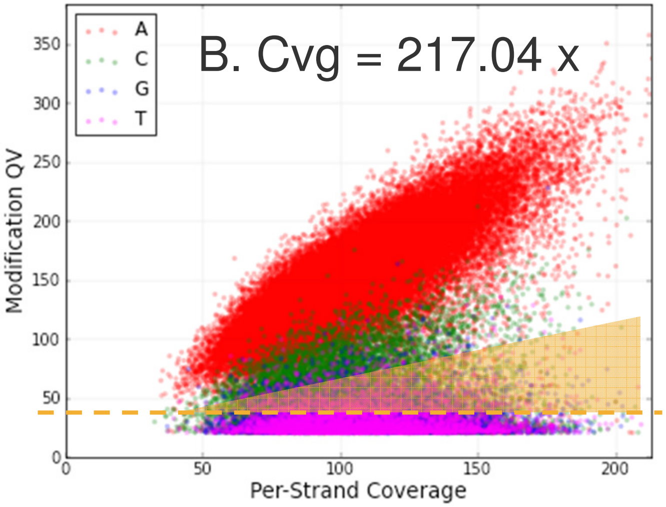
Apply knowledge about how bacterial restriction / modification systems work

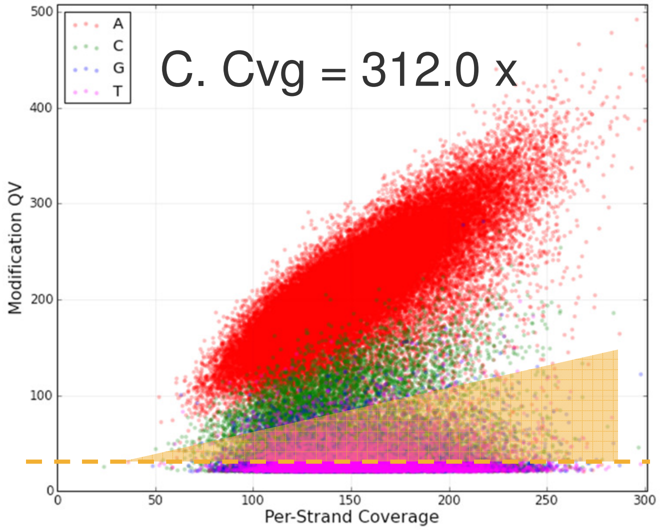
- Modification of motifs is generally near 100% in bacterial systems.
- Motifs are often palindromic, and modified on both strands in a reverse complimentary manner

Make use of public databases on methylation systems

Principle 1: Limit Motif Finding to High Confidence Hits







Example: *Thermoplasma acidophilum*

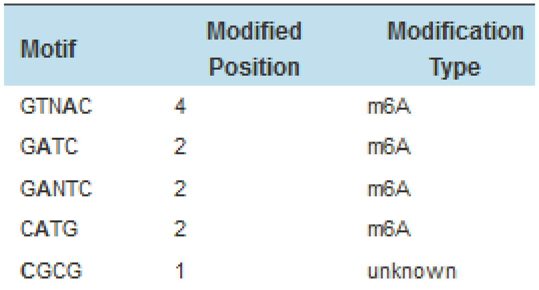
The higher your coverage, the more false positives you include when you use the default minimum modification QV value of 30.

If you ask a computer to find a pattern in noise, it often will.

Motif Finding Results Will Vary Widely With Coverage Using the Default Min Mod QV = 30

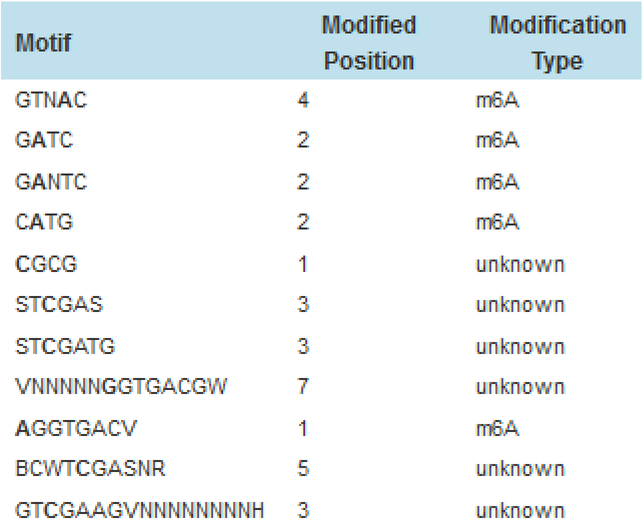
Thermoplasma acidophilum

A. Coverage = 114.83 x



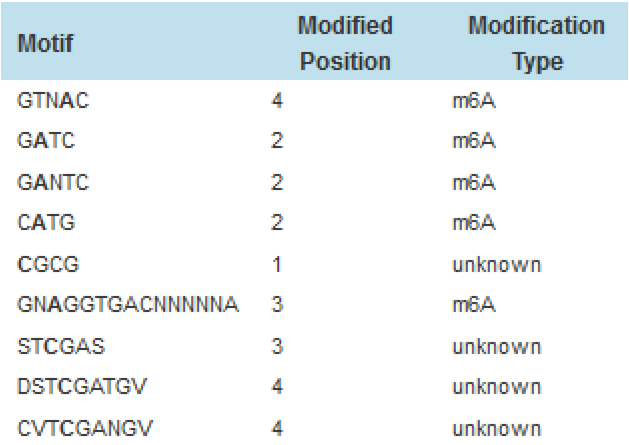
Motif	Modified Position	Modification Type
GTNAC	4	m6A
GATC	2	m6A
GANTC	2	m6A
CATG	2	m6A
CGCG	1	unknown

B. Coverage = 217.04 x



Motif	Modified Position	Modification Type
GTNAC	4	m6A
GATC	2	m6A
GANTC	2	m6A
CATG	2	m6A
CGCG	1	unknown
STCGAS	3	unknown
STCGATG	3	unknown
VNNNNNGGTGACGW	7	unknown
AGGTGACV	1	m6A
BCWTCGASNR	5	unknown
GTCGAAGVNNNNNNNNH	3	unknown

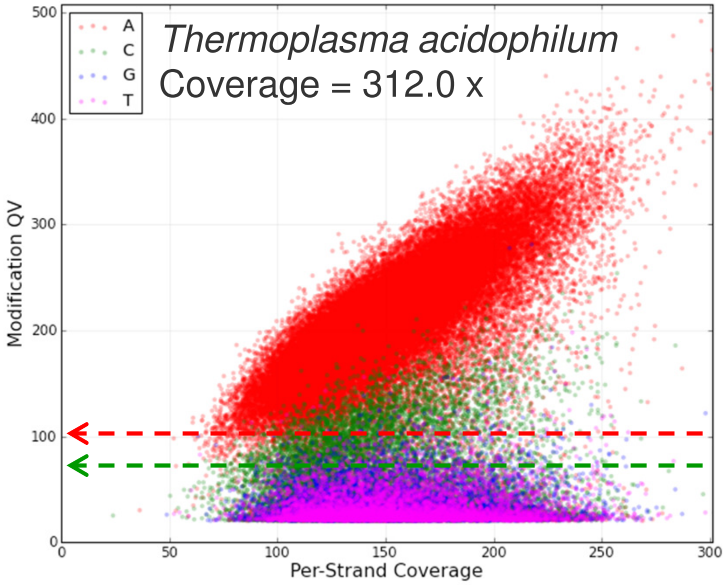
C. Coverage = 312.0 x



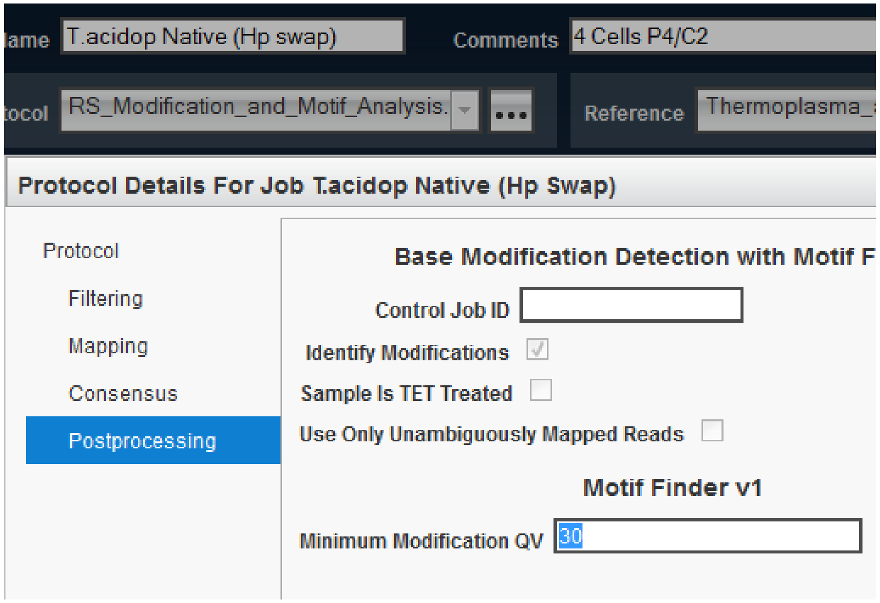
Motif	Modified Position	Modification Type
GTNAC	4	m6A
GATC	2	m6A
GANTC	2	m6A
CATG	2	m6A
CGCG	1	unknown
GNAGGTGACNNNNNA	3	m6A
STCGAS	3	unknown
DSTCGATGV	4	unknown
CVTCGANGV	4	unknown

Case A - which pairs the recommended coverage with the default min QV setting, gives the cleanest result.

Adjusting the Minimum Modification QV Setting Often Improves the Clarity of Results



Estimate the correct min QV using the Modification QV vs Per-Strand Coverage plot.



lame T.acidop Native (Hp swap) Comments 4 Cells P4/C2

tocol RS_Modification_and_Motif_Analysis. Reference Thermoplasma_

Protocol Details For Job T.acidop Native (Hp Swap)

Base Modification Detection with Motif F

Control Job ID

Identify Modifications ☑

Sample Is TET Treated □

Use Only Unambiguously Mapped Reads □

Motif Finder v1

Minimum Modification QV 30

You may have to balance between reducing false positive data points and retaining weak but true ‘C’ signals.

Try multiple settings!

Motif Maker

- Tool that allows you to rapidly rerun just the motif-finding part of the SMRT Portal protocol using different minimum-modification QV settings

- Command-line java tool that runs on any platform, as long as java is installed

- Detailed information on usage can be found here:

  https://github.com/PacificBiosciences/MotifMaker

- Let’s Try it Ourselves:

```
$ wget https://github.com/PacificBiosciences/MotifMaker/archive/master.zip

$ unzip master.zip

$ java -jar MotifMaker-master/target/motif-maker-0.1.one-jar.jar find -f
Thermoplasma_acidophilum_DSM1728.fasta -g modifications.gff.gz -o motifsQV60.csv –
m 60
```

How Does Motif Finding Work?

- Algorithm searches for high scoring motifs in the 41-base sequence contexts written to the modifications.gff file, using only entries where the modification QV is above the specified cutoff.
- Progressively longer motifs are tested, until no more commonalities are found, or the motif length limit is reached.

```
Score = nDetected / nGenome * (sum of log-pvalue of detected motifs)
```

- Since longer motifs will have fewer sites in the genome, over-constraining a motif can inflate the score above the reporting threshold for a motif where the kinetic signal may be weak (low coverage situations, 5mC).

```
Thermoplasma acidophilum  example with coverage = 312.0 and min QV = 30
DSTCGATGV         63  /    218  *   792 =   229.0    REPORTED
TCGA             765  / 12,464  * 2,229 =   136.8    NOT REPORTED
```

Take home : Sets of bizarre motifs with a common core often collectively hint at weakly detected true motifs.

Motif Finding Results with a stringent min QV = 100

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 100

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.992	4,789	4,826	219.76	150.22
GATC	1	m6A	0.988	25,701	26,020	217.07	149.32
CATG	1	m6A	0.986	15,956	16,180	222.00	152.65
GANTC	1	m6A	0.977	9,265	9,482	214.28	149.35
CGCG	0	m4C	0.398	1,737	4,366	128.77	161.27

Reporting of the easily detected m6A motifs is essentially unchanged by applying the stringent min QV setting of 100.

Motif Finding Results with a stringent min QV = 100

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 100

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.992	4,789	4,826	219.76	150.22
GATC	1	m6A	0.988	25,701	26,020	217.07	149.32
CATG	1	m6A	0.986	15,956	16,180	222.00	152.65
GANTC	1	m6A	0.977	9,265	9,482	214.28	149.35
CGCG	0	m4C	0.398	1,737	4,366	128.77	161.27

The m4C motif, however, is much less robustly detected, since the score of most **C**GCG motifs is < 100.

PACIFIC BIOSCIENCES®

Motif Finding Results with a stringent min QV = 100

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 100

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.992	4,789	4,826	219.76	150.22
GATC	1	m6A	0.988	25,701	26,020	217.07	149.32
CATG	1	m6A	0.986	15,956	16,180	222.00	152.65
GANTC	1	m6A	0.977	9,265	9,482	214.28	149.35
CGCG	0	m4C	0.398	1,737	4,366	128.77	161.27

This odd motif disappears with the more stringent setting. Can anyone explain the source of the motif?

Motif Finding Results with a stringent min QV = 100

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 100

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.992	4,789	4,826	219.76	150.22
GATC	1	m6A	0.988	25,701	26,020	217.07	149.32
CATG	1	m6A	0.986	15,956	16,180	222.00	152.65
GANTC	1	m6A	0.977	9,265	9,482	214.28	149.35
CGCG	0	m4C	0.398	1,737	4,366	128.77	161.27

Principle 2: Apply what you know about base mod signals. This is part of the m6A footprint of GTNAC – a secondary signal occurs 5 bases upstream of m6A.

Motif Finding Results with a stringent min QV = 100

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 100

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.992	4,789	4,826	219.76	150.22
GATC	1	m6A	0.988	25,701	26,020	217.07	149.32
CATG	1	m6A	0.986	15,956	16,180	222.00	152.65
GANTC	1	m6A	0.977	9,265	9,482	214.28	149.35
CGCG	0	m4C	0.398	1,737	4,366	128.77	161.27

Finally, these over-constrained motifs have disappeared with the more stringent setting.

Motif Finding Results with Min QV = 60

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 60

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.999	4,823	4,826	218.81	149.96
GATC	1	m6A	0.998	25,968	26,020	215.73	149.00
GANTC	1	m6A	0.996	9,440	9,482	211.88	148.94
CATG	1	m6A	0.995	16,100	16,180	220.79	152.44
CGCG	0	m4C	0.782	3,415	4,366	105.34	145.71
VNNNNNDNVSTCGAG	11	modified_base	0.187	52	278	77.69	176.08
AGGTGACV	0	m6A	0.159	10	63	100.80	151.70

These motifs with the strongest signals are again largely unchanged with the more moderate min QV setting. Note the better retention of the m4C motif CGCG.

Motif Finding Results with Min QV = 60

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 60

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.999	4,823	4,826	218.81	149.96
GATC	1	m6A	0.998	25,968	26,020	215.73	149.00
GANTC	1	m6A	0.996	9,440	9,482	211.88	148.94
CATG	1	m6A	0.995	16,100	16,180	220.79	152.44
CGCG	0	m4C	0.782	3,415	4,366	105.34	145.71
VNNNNNDNVSTCGAG	11	modified_base	0.187	52	278	77.69	176.08
AGGTGACV	0	m6A	0.159	10	63	100.80	151.70

We again see a motif that is really just part of the GTNAC footprint.

PACIFIC BIOSCIENCES®

Motif Finding Results with Min QV = 60

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 60

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.999	4,823	4,826	218.81	149.96
GATC	1	m6A	0.998	25,968	26,020	215.73	149.00
GANTC	1	m6A	0.996	9,440	9,482	211.88	148.94
CATG	1	m6A	0.995	16,100	16,180	220.79	152.44
CGCG	0	m4C	0.782	3,415	4,366	105.34	145.71
VNNNNNDNVSTCGAG	11	modified_base	0.187	52	278	77.69	176.08
AGGTGACV	0	m6A	0.159	10	63	100.80	151.70

Any guess on what's up with these bizarre motifs?

PACIFIC BIOSCIENCES®

Motif Finding Results with Min QV = 60

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 60

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.999	4,823	4,826	218.81	149.96
GATC	1	m6A	0.998	25,968	26,020	215.73	149.00
GANTC	1	m6A	0.996	9,440	9,482	211.88	148.94
CATG	1	m6A	0.995	16,100	16,180	220.79	152.44
CGCG	0	m4C	0.782	3,415	4,366	105.34	145.71
VNNNNNDNVSTCGAG	11	modified_base	0.187	52	278	77.69	176.08
AGGTGACV	0	m6A	0.159	10	63	100.80	151.70

Principle 2: Apply what you know about base mod signals and motif finding. These low-scoring motifs are weak m5C signals. Because of the weakness signal, the only the over-constrained motifs score highly enough to be reported.

Motif Finding Results with Min QV = 60

default minQV = 30

motif	position*	type	fraction	nDetected	nGenome	score	coverage
GTNAC	4	m6A	100.00	4,826	4,826	218.7	149.90
GATC	2	m6A	99.94	26,004	26,020	215.5	149.00
GANTC	2	m6A	99.89	9,472	9,482	211.3	148.90
CATG	2	m6A	99.77	16,142	16,180	220.3	152.40
CGCG	1	unknown	91.48	3,994	4,366	97.1	142.30
GNAGGTGACNNNNNA	3	m6A	83.33	5	6	91	170.80
STCGAS	3	unknown	38.32	410	1,070	50.4	158.00
DSTCGATGV	4	unknown	28.90	63	218	43.5	159.50
CVTCGANGV	4	unknown	28.02	188	671	48.2	158.20

min QV = 60

motif	position	type	fraction	nDetected	nGenome	score	coverage
GTNAC	3	m6A	0.999	4,823	4,826	218.81	149.96
GATC	1	m6A	0.998	25,968	26,020	215.73	149.00
GANTC	1	m6A	0.996	9,440	9,482	211.88	148.94
CATG	1	m6A	0.995	16,100	16,180	220.79	152.44
CGCG	0	m4C	0.782	3,415	4,366	105.34	145.71
VNNNNNDNVSTCGAG	11	modified_base	0.187	52	278	77.69	176.08
AGGTGACV	0	m6A	0.159	10	63	100.80	151.70

Principle 3: Apply what we know about RM systems. The true motif is almost certainly the palindromic motif TCGA, hiding in the middle of all the various unaccounted for motifs reported at different min QV settings.

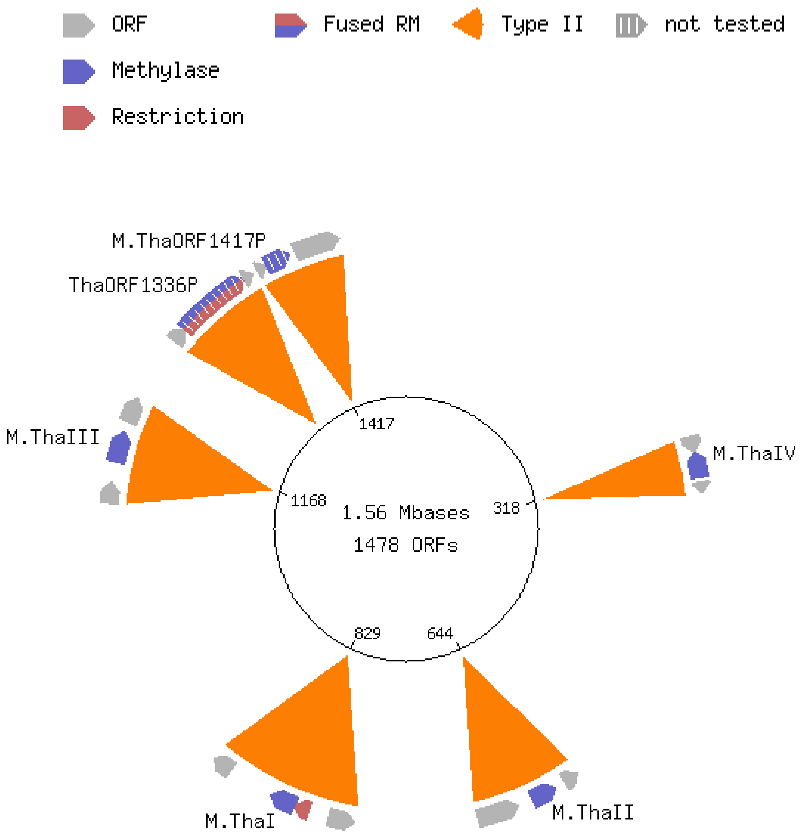
REBASE Genomics

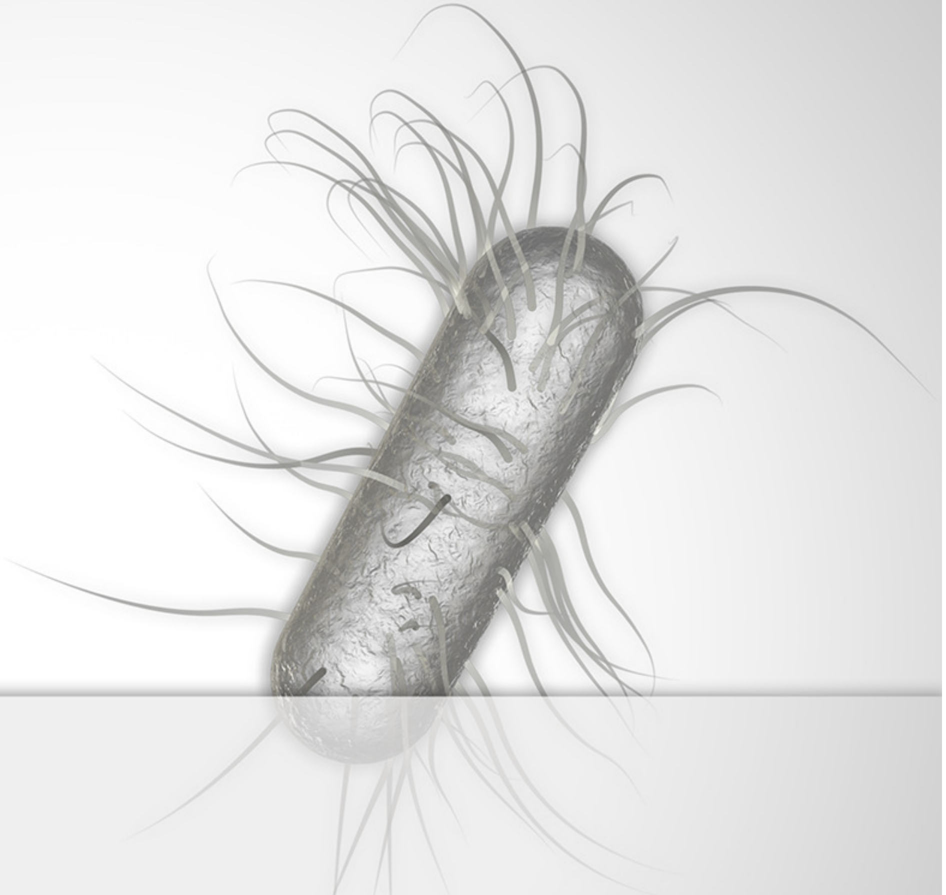
Principle 4 : Make use of publicly available databases on RM systems.
http://tools.neb.com/~vincze/genomes/

Type	Gene	Name	Predicted
			Rec Seq
II	**M**	**M.ThaIV**	**CATG**
II	M	M.ThaII	GATC
II	**R**	**ThaI**	**CGCG**
II	**M**	**M.ThaI**	**CGCG**
II	M	M.ThaIII	GANTC
II	**RM**	**ThaORF1336P**	TCGA /?GTNAC
II	M	M.ThaORF1417P	TCGA / GTNAC

The RM gene prediction concords nicely with our results.

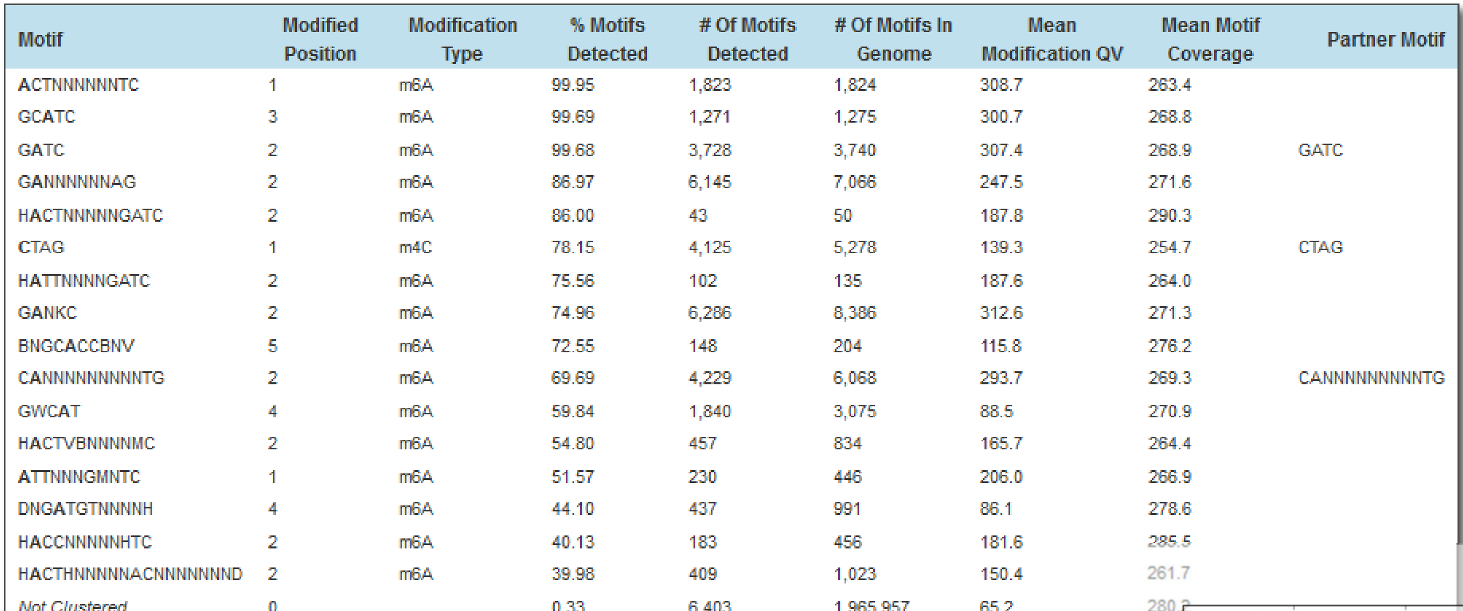
http://tools.neb.com/~vincze/genomes/summary.php?genome_id=225



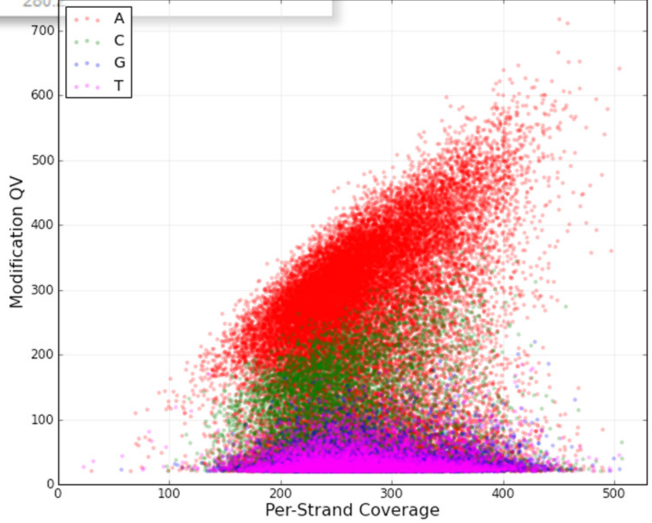


M. bovis: Using R to Refine Motif Results

Using R to Refine Base Modification Data : *Mycoplasma bovis*



Motif	Modified Position	Modification Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
ACTNNNNNNTC	1	m6A	99.95	1,823	1,824	308.7	263.4	
GCATC	3	m6A	99.69	1,271	1,275	300.7	268.8	
GATC	2	m6A	99.68	3,728	3,740	307.4	268.9	GATC
GANNNNNNAG	2	m6A	86.97	6,145	7,066	247.5	271.6	
HACTNNNNNGATC	2	m6A	86.00	43	50	187.8	290.3	
CTAG	1	m4C	78.15	4,125	5,278	139.3	254.7	CTAG
HATTNNNNGATC	2	m6A	75.56	102	135	187.6	264.0	
GANKC	2	m6A	74.96	6,286	8,386	312.6	271.3	
BNGCACCBNV	5	m6A	72.55	148	204	115.8	276.2	
CANNNNNNNNNTG	2	m6A	69.69	4,229	6,068	293.7	269.3	CANNNNNNNNNTG
GWCAT	4	m6A	59.84	1,840	3,075	88.5	270.9	
HACTVBNNNNMC	2	m6A	54.80	457	834	165.7	264.4	
ATTNNNGMNTC	1	m6A	51.57	230	446	206.0	266.9	
DNGATGTNNNNH	4	m6A	44.10	437	991	86.1	278.6	
HACCNNNNNHTC	2	m6A	40.13	183	456	181.6	285.5	
HACTHNNNNNACNNNNNNND	2	m6A	39.98	409	1,023	150.4	261.7	
Not Clustered	0		0.33	6,403	1,965,957	65.2	280.2	



Here are initial motif finding results from 5 SMRT Cells using the default minimum modification QV setting of 30.

Let's import the data into R and learn a second way to refine our results.

PACIFIC BIOSCIENCES®

Refining Base Modification Results With R

- We have written a number of functions to facilitate more in-depth or custom analysis in R:
  - Do more nuanced, custom filtering of hits by score and coverage
  - Annotate any motif of interest and refine the SMRTPortal results
  - Plot the score vs. coverage distribution by base
  - Examine the distribution of score, coverage, IPD Ratio or other factors for any motif of interest, both modified and unmodified
  - Visualize your results using circos
- Example data and R functions can be found online:
  - https://github.com/PacificBiosciences/Bioinformatics-Training/
  - http://pacb.com/bmd/ (basemod data sets at PacBio)
  - https://github.com/PacificBiosciences/R-kinetics (github R Kinetics package)

Launching RStudio

- Either RGui (PC or Mac version 2.15.0) or RStudio (which interfaces with a Linux or cloud installation) can be used. For this tutorial we will use RStudio
- **RStudio** : http://ec2-54-85-54-242.compute-1.amazonaws.com:8787/
- **ssh**: ec2-54-85-54-242.compute-1.amazonaws.com

Use PuTTY to ssh into the AMI with your user number and the password ‘Pacbio150k’

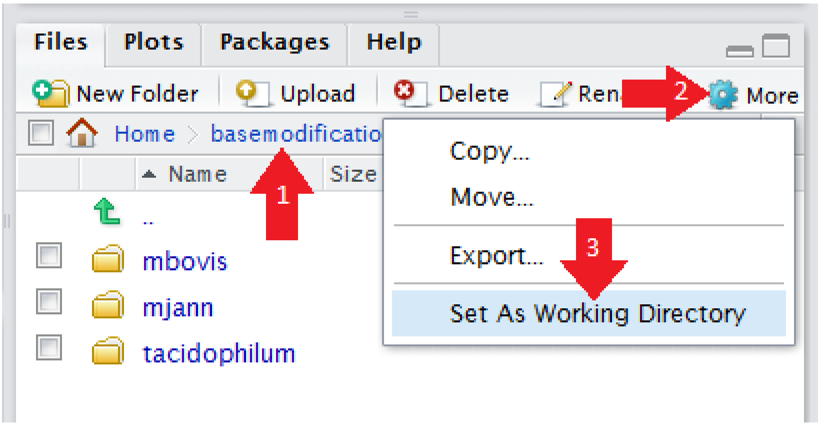
Copy the tutorial materials into your home directory with the following command:

```
$ cp -r /training/basemodification ~
```

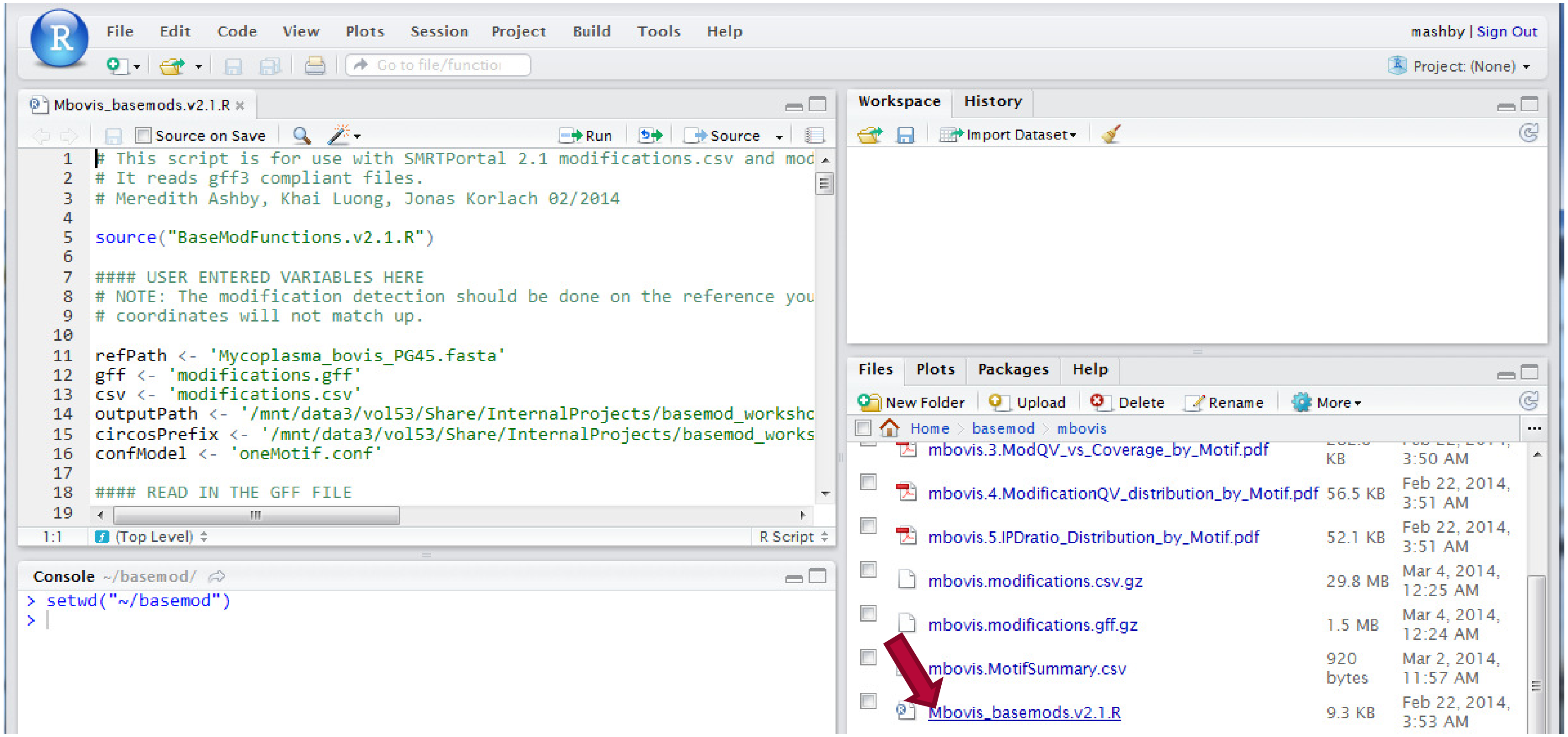
Open RStudio in your browser with the above link and log in.

In the bottom right quadrant, move to the basemodification folder.

In the ‘More’ pull-down menu, select ‘Set As Working Directory’.



Continue the Tutorial in R by Opening Mbovis_basemods.v2.1.R



```
# This script is for use with SMRTPortal 2.1 modifications.csv and moc
# It reads gff3 compliant files.
# Meredith Ashby, Khai Luong, Jonas Korlach 02/2014

source("BaseModFunctions.v2.1.R")

#### USER ENTERED VARIABLES HERE
# NOTE: The modification detection should be done on the reference you
# coordinates will not match up.

refPath <- 'Mycoplasma_bovis_PG45.fasta'
gff <- 'modifications.gff'
csv <- 'modifications.csv'
outputPath <- '/mnt/data3/vol53/Share/InternalProjects/basemod_workshc
circosPrefix <- '/mnt/data3/vol53/Share/InternalProjects/basemod_works
confModel <- 'oneMotif.conf'

#### READ IN THE GFF FILE
```

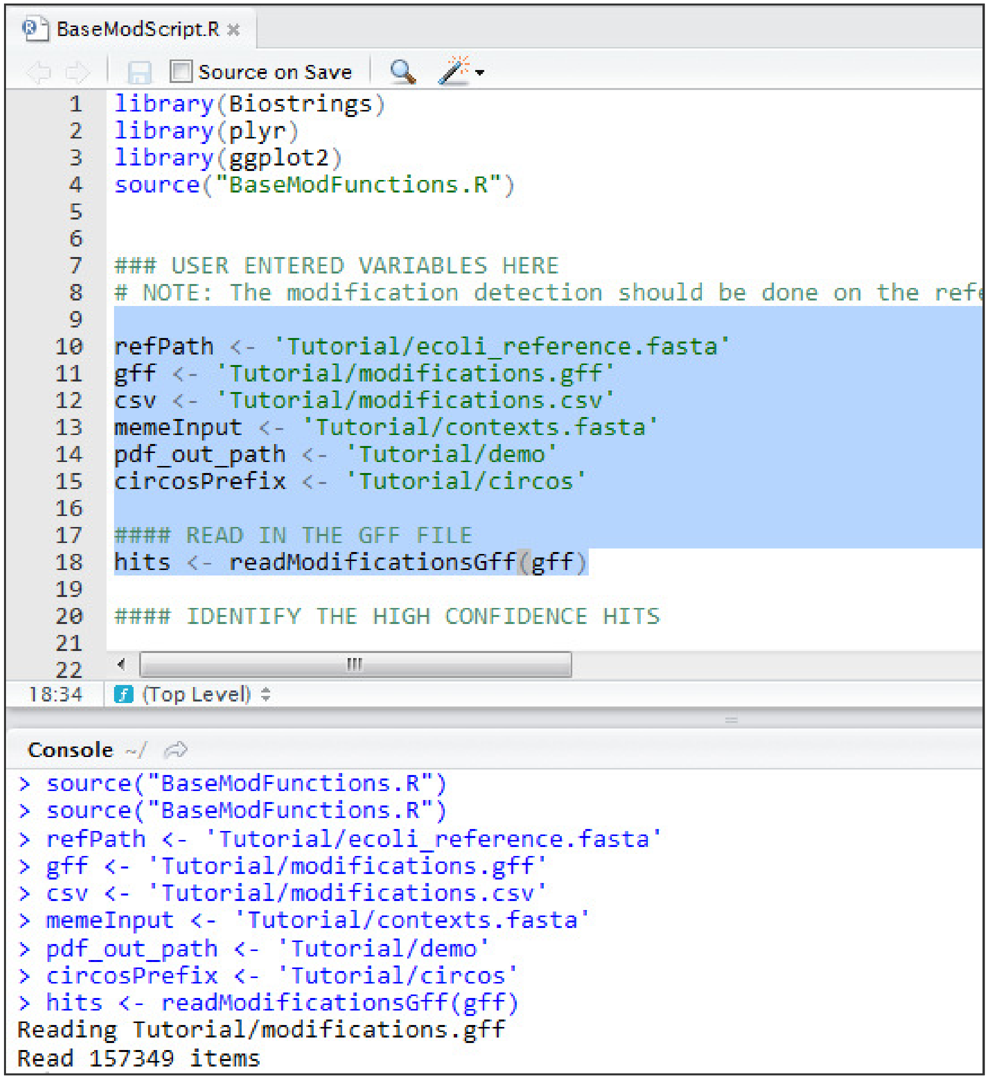
```
> setwd("~/basemod")
> 
```

mbovis.3.ModQV_vs_Coverage_by_Motif.pdf	KB	3:50 AM
mbovis.4.ModificationQV_distribution_by_Motif.pdf	56.5 KB	Feb 22, 2014, 3:51 AM
mbovis.5.IPDratio_Distribution_by_Motif.pdf	52.1 KB	Feb 22, 2014, 3:51 AM
mbovis.modifications.csv.gz	29.8 MB	Mar 4, 2014, 12:25 AM
mbovis.modifications.gff.gz	1.5 MB	Mar 4, 2014, 12:24 AM
mbovis.MotifSummary.csv	920 bytes	Mar 2, 2014, 11:57 AM
Mbovis_basemods.v2.1.R	9.3 KB	Feb 22, 2014, 3:53 AM

Open the mbovis folder and single click Mbovis_basemods.v2.1.R to open it in the console.

If you use this script later as a template for your own analyses, you will have to edit the input and output paths to match the directory on your own server.

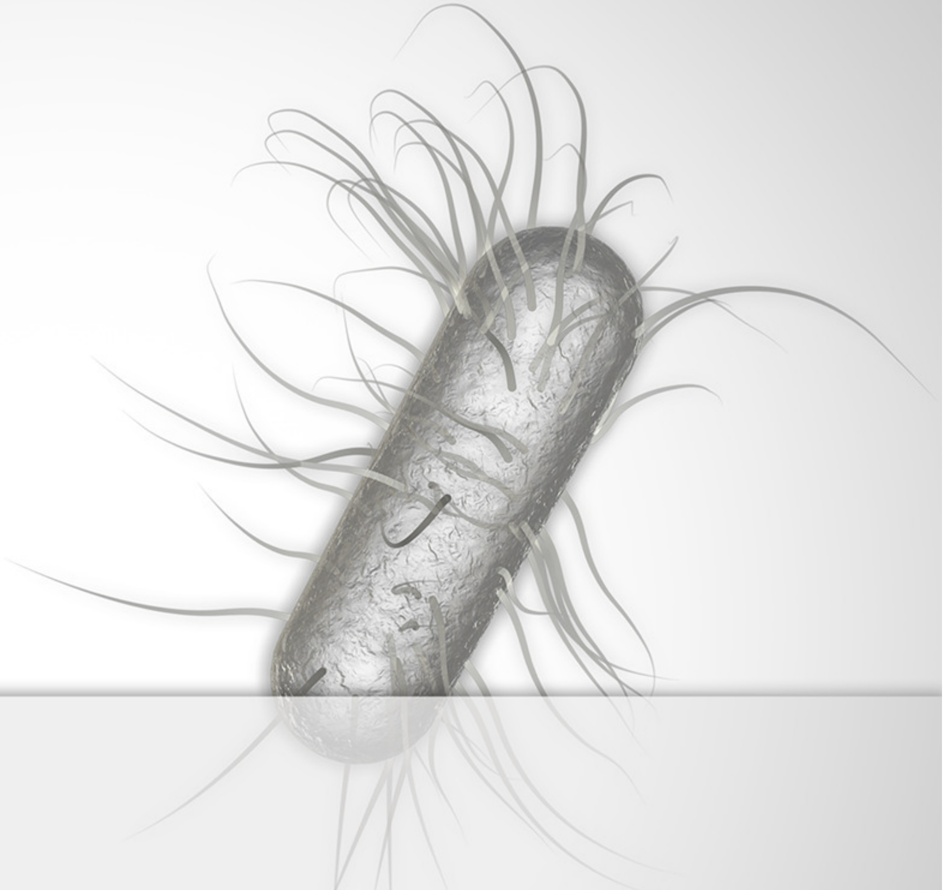
Execute Blocks of Code with ‘Ctrl + Enter’



```
library(Biostrings)
library(plyr)
library(ggplot2)
source("BaseModFunctions.R")


### USER ENTERED VARIABLES HERE
# NOTE: The modification detection should be done on the refe

refPath <- 'Tutorial/ecoli_reference.fasta'
gff <- 'Tutorial/modifications.gff'
csv <- 'Tutorial/modifications.csv'
memeInput <- 'Tutorial/contexts.fasta'
pdf_out_path <- 'Tutorial/demo'
circosPrefix <- 'Tutorial/circos'

#### READ IN THE GFF FILE
hits <- readModificationsGff(gff)

#### IDENTIFY THE HIGH CONFIDENCE HITS


```

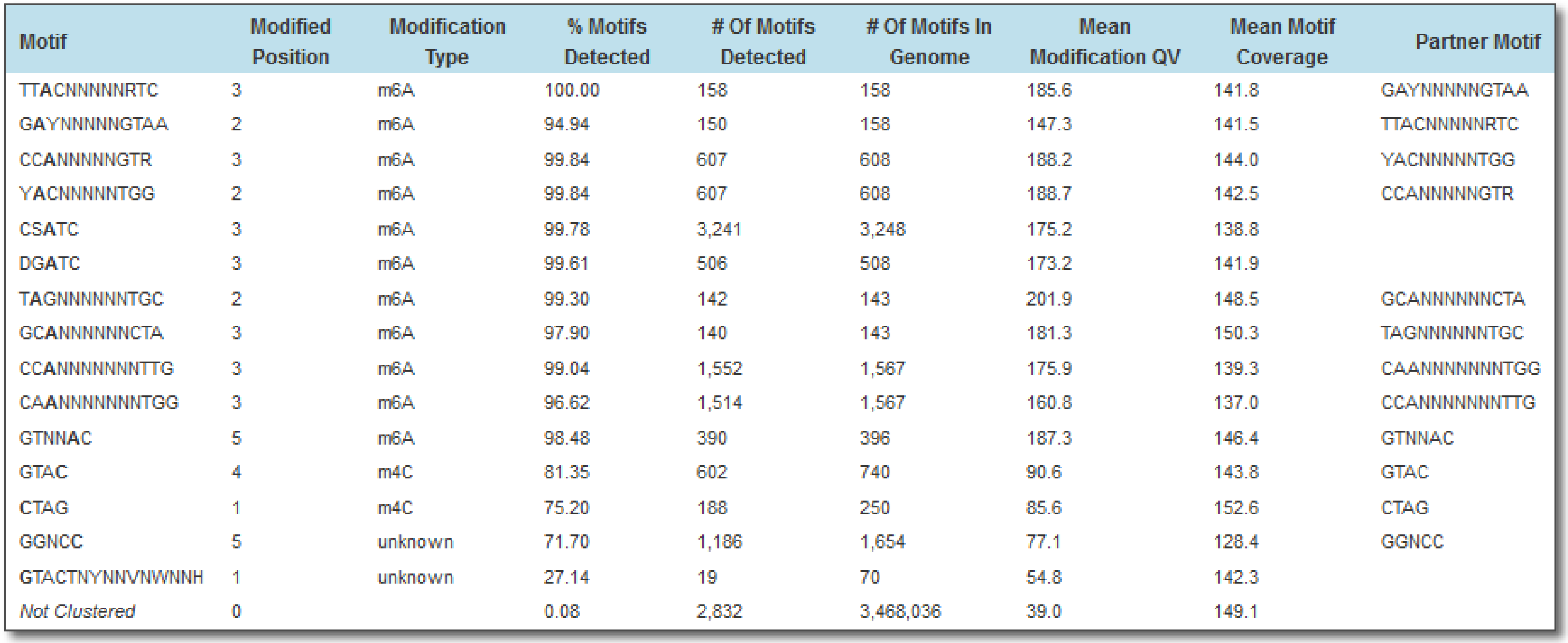
```
> source("BaseModFunctions.R")
> source("BaseModFunctions.R")
> refPath <- 'Tutorial/ecoli_reference.fasta'
> gff <- 'Tutorial/modifications.gff'
> csv <- 'Tutorial/modifications.csv'
> memeInput <- 'Tutorial/contexts.fasta'
> pdf_out_path <- 'Tutorial/demo'
> circosPrefix <- 'Tutorial/circos'
> hits <- readModificationsGff(gff)
Reading Tutorial/modifications.gff
Read 157349 items
```

- To being, highlight the library commands, the block of path variables and the command to read in the gff file
- Hit ‘Ctrl + Enter’ to run all the highlighted lines in the Console
- The gff file will be read into a data.frame called hits
- To see how the function ‘readModificationsGFF’ or any of the other functions used here works, you can open up BaseModFunctions.R
- Continue in this way through to the end of the *M. bovis* example.

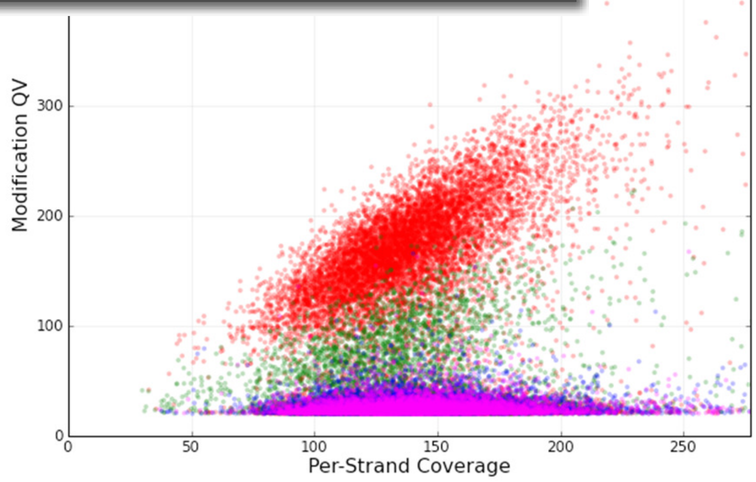


M. Jannaschii : Applying What We’ve Learned

Apply What We've Learned : *Methanocaldococcus jannaschii*



Motif	Modified Position	Modification Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
TTACNNNNNRTC	3	m6A	100.00	158	158	185.6	141.8	GAYNNNNNGTAA
GAYNNNNNGTAA	2	m6A	94.94	150	158	147.3	141.5	TTACNNNNNRTC
CCANNNNNGTR	3	m6A	99.84	607	608	188.2	144.0	YACNNNNNTGG
YACNNNNNTGG	2	m6A	99.84	607	608	188.7	142.5	CCANNNNNGTR
CSATC	3	m6A	99.78	3,241	3,248	175.2	138.8	
DGATC	3	m6A	99.61	506	508	173.2	141.9	
TAGNNNNNNTGC	2	m6A	99.30	142	143	201.9	148.5	GCANNNNNNCTA
GCANNNNNNCTA	3	m6A	97.90	140	143	181.3	150.3	TAGNNNNNNTGC
CCANNNNNNNTTG	3	m6A	99.04	1,552	1,567	175.9	139.3	CAANNNNNNNTGG
CAANNNNNNNTGG	3	m6A	96.62	1,514	1,567	160.8	137.0	CCANNNNNNNTTG
GTNNAC	5	m6A	98.48	390	396	187.3	146.4	GTNNAC
GTAC	4	m4C	81.35	602	740	90.6	143.8	GTAC
CTAG	1	m4C	75.20	188	250	85.6	152.6	CTAG
GGNCC	5	unknown	71.70	1,186	1,654	77.1	128.4	GGNCC
GTACTNYNNVNWNNH	1	unknown	27.14	19	70	54.8	142.3	
Not Clustered	0		0.08	2,832	3,468,036	39.0	149.1	



Please use the remaining time to use what you have learned to generate a refined motif list for *Methanocaldococcus jannaschii*.

Feel free to use whatever tools you prefer, and to work either solo or with a partner.

HINT: Be skeptical of triply degenerate bases.

PACIFIC
BIOSCIENCES®

Methanocaldococcus jannaschii : High Confidence Motif List

Motif	Pos	Type	Fraction	nDetected	nGenome	Score	Cvg	Partner Motif
CAANNNNNNNTGG	3	6mA	96.6	1514	1567	160.8	137.0	CCANNNNNNNTTG
CCANNNNNNNTTG	3	6mA	99.0	1552	1567	175.9	139.3	CAANNNNNNNTGG
GATC	*3*	*6mA*	*99.6*	*506*	*508*	*173.2*	*141.9*	*GATC*
CCATC	*3*	*6mA*	*99.8*	*3241*	*3248*	*175.2*	*138.8*	
CTAG	1	4mC	75.2	188	250	85.6	152.6	CTAG
GAYNNNNNGTAA	2	6mA	94.9	150	158	147.3	141.5	TTACNNNNNRTC
TTACNNNNNRTC	3	6mA	100.0	158	158	185.6	141.8	GAYNNNNNGTAA
GCANNNNNNCTA	3	6mA	97.9	140	143	181.3	150.3	TAGNNNNNNTGC
TAGNNNNNNTGC	2	6mA	99.3	142	143	201.9	148.5	GCANNNNNNCTA
GGNCC	5	4mC	71.7	1186	1654	77.1	128.4	GGNCC
GTAC	4	4mC	81.4	602	740	90.6	143.8	GTAC
GTNNAC	5	6mA	98.5	390	396	187.3	146.4	GTNNAC
CCANNNNNGTR	3	6mA	99.8	607	608	188.2	144.0	YACNNNNNTGG
YACNNNNNTGG	2	6mA	99.8	607	608	188.7	142.5	CCANNNNNGTR

Two Motifs must be untangled to arrive at the correct answer:

Motif finder calls GATC as DGATC due to overlap with CCATC
Motif finder calls CCATC as CSATC due to overlap with GATC

Getting Started With R on Your Own System

- Installing R on your PC
  - http://cran.r-project.org/bin/windows/base/

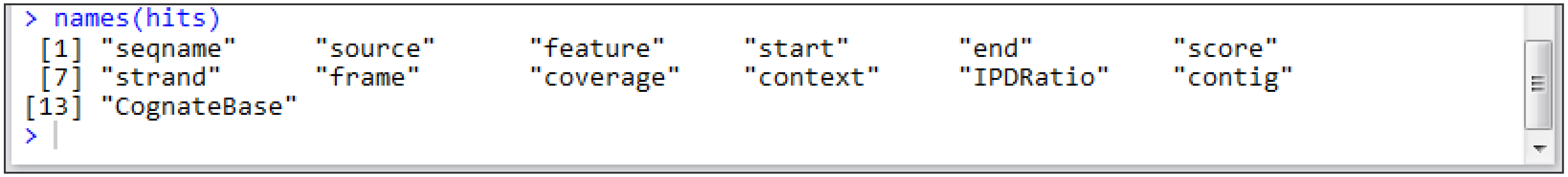
- Online help with R
  - http://www.r-project.org/
  - http://www.ats.ucla.edu/stat/R/faq/
  - http://stat.ethz.ch/R-manual/R-devel/doc/manual/R-lang.html
  - http://had.co.nz/ggplot2/

- Outside the scope of this tutorial, but highly recommended, is becoming comfortable using R in the Unix environment. All the commands are the same, but you will need to install putty.exe and learn to use a unix text editor (emacs is highly recommended).

Getting Comfortable With R

- A useful reference for getting started with R can be found here:
  - http://cran.r-project.org/doc/manuals/R-intro.pdf

- For now, try these commands, which are handy for examining any dataframe:
  - names(hits)
  - dim(hits)
  - head(hits)
  - table(hits$source, hits$feature)
  - levels(factor(hits$feature))
  - ls()



```
> names(hits)
 [1] "seqname"     "source"      "feature"     "start"       "end"         "score"
 [7] "strand"      "frame"       "coverage"    "context"     "IPDRatio"    "contig"
[13] "CognateBase"
> 
```