# Meraculous
# *De Novo* Assembly of the *Ariolimax dolichophallus* Genome

Charles Cole, Jake Houser, Kyle McGovern, and Jennie Richardson

# The Meraculous Hash

## What critics are saying:

Seems too complicated? We will expand on this commentary as we work through their code.

**[Brief description of exactly what the paper said]**

[To be continued]

http://www.homolog.us/blogs/blog/2012/10/02/perfect-hash-algorithm-of-meraculous-assembler/

*lightweight hash*
http://plosone.org/article/info:doi/10.1371/journal.pone.0023501#article1.body1.sec2.sec7.p1

Did authors came up with this data structure completely on their own - is it a complete novelty ? If that is the case it should be described in more detail. Otherwise they should provide a reference to the publication were the idea was drawn from.

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023501

# The Meraculous Hash

Stated Goals

- Don’t store keys
- Perfect, static hash
- Each value is [ACTG][ACTG] representing the unique forward and backwards extensions
- Only hash kmers that have U-U extensions

# The Meraculous Hash

## The Implementation

First, the hash must be "primed" as follows: (we assume there are hash functions $h_0 \ldots h_n$ already defined).

1. Initialize hash depth d to 0, write all keys to file $F_d$.
2. For all keys in file $F_d$, evaluate the hash function $h_d$ and update a "primer object" $P_d$ to keep track of which hash values occur multiple times at hash depth d (i.e. the keys collide under the hash function $h_d$).
3. Write all colliding keys to file $F_{d+1}$ ; increment hash depth d.
4. Repeat steps 1,2 until the number of colliding keys is 0.

# The Meraculous Hash

- Irrelevant?
- Current version of Meraculous is 2.0.5, 4 years newer than the paper
- No mention of “novel lightweight hash” in documentation or website
- What does the source code say?
- No mention of any of the hash functionality they described in the original paper

# The Meraculous Hash

- Distributed hash tables
- Multithreaded generation of hash table files using boost
- Output a number of UFX.N files where N is a 3-4 nucleotide string

```
       725 May  6 19:31 UFX.ATTC.err
6169243925 May  6 19:31 UFX.CAAT
       727 May  6 19:31 UFX.CAAT.err
5489404895 May  6 19:17 UFX.CAGT
       724 May  6 19:17 UFX.CAGT.err
4988678370 May  6 19:15 UFX.CCAG
       725 May  6 19:15 UFX.CCAG.err
4805340995 May  6 19:15 UFX.CCTT
       724 May  6 19:15 UFX.CCTT.err
```

# The Meraculous Hash

- 73 GB of UFX files from our completed run
- What do these files look like?
- Each file contains lines of:
  - [kmer]  [ACTG][ACTG]

```
AAAAAAAAAATGTAGAACAGATATAAACATGTATGTCACGTACCGAGCTAAGAGGGAAACAA   TA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAATATTTTTTTTTTTTTTAAT   AX
AAAAAAAAATATTTCATAATATCTAATCTGGACCTTTCTTGAAACAATTTGAAGCTGTGTAC   TT
AAAAAAAAAAAGACAATGAGACAGCGAAAATAAGCAAGAAAAATTAGGCAGAGAAAATAAGG   FG
AAAAAAAAAAAAAACCTTTAACATAATGGATTATTTGTCATTGGCAGTTGTCTTGCAAGTTT   TT
AAAAAAAAAAACCCATTCCTCAGCTCCTGACCAAATAAAAAACCCCATTCCTTAGGGTTCTT   XX
AAAAAAAAAAAAAACCAGAAAATAAATATAATGCATTGTTTATGCAAATATAATAAGGTTAC   AT
AAAAAAAAAAAACCCCGACAACAGATTATCGATGTTTACCATAACGGTCATATCTGTCAGAT   AT
AAAAAAAATTAACAGCACTGCAAAAAGTTTGGCACAAAAATAAACACGTTGTAGACACTGC   GC
```

# The Meraculous Hash

- What about memory usage?
- We still have to look up kmers in the hash to traverse the deBruijn graph
- Memory usage is worse than the original implementation because we hash ALL extensions, not just U-U
- Packed value storage to reduce kmer footprint
- Each kmer is divided into chunks of 4-nucleotide blocks and then converted to an int that uniquely maps that block
- Hash stores the packed kmer, value

# User Experience

Update:

- SGE memory issue resolved
- Modified program so that minimum coverage is not needed
- Program can now produce contigs
- Program fails at the bubble-popping step

# Memory issues

- Program was dieing during the kmer counting stage due to uncompleted jobs
- testing of the code revealed that qsub can’t be run with the “-w e” option while specifying the mem_free resource

```
# constant global options
opts_constant="-v $env_vars -cwd -r n -b n -S /bin/bash -w e -j y -q all.q "
qsub_opts+=$opts_constant


##  The command line options are supplied by Meraculous, but you can control if
submitter

project=
name=
out_dir=
n_tasks=
n_slots=
ram_req=
wclk_req=

while getopts e:n:o:p:r:s:t:w: opt; do
  case $opt in
  n)  # name for the job
      name=$OPTARG
      qsub_opts+=" -N $name"
      ;;
  o)  # default stdout directory - this is where uncaptured stdout plus any clus
      out_dir=$OPTARG
      qsub_opts+=" -o $out_dir"
      ;;
  e)  # default stderr directory - this is where uncaptured stderr will go
      out_dir=$OPTARG
      qsub_opts+=" -e $out_dir"
      ;;
  p)  # project name
      project=$OPTARG
      qsub_opts+=" -P $project"
      ;;
  r)  # cluster ram request (will be applied *per slot* if n_slots is set
      ram_req=$OPTARG
      qsub_opts+=" -l h_vmem=$ram_req"
      ;;
  s)  # number of slots requested per task
```

```
# constant global options
opts_constant="-v $env_vars -cwd -r n -b n -S /bin/bash -w e -j y -q all.q "
qsub_opts+=$opts_constant


##  The command line options are supplied by Meraculous, but you can control if and
submitter

project=
name=
out_dir=
n_tasks=
n_slots=
ram_req=
wclk_req=

while getopts e:n:o:p:r:s:t:w: opt; do
  case $opt in
  n)  # name for the job
      name=$OPTARG
      qsub_opts+=" -N $name"
      ;;
  o)  # default stdout directory - this is where uncaptured stdout plus any cluster
      out_dir=$OPTARG
      qsub_opts+=" -o $out_dir"
      ;;
  e)  # default stderr directory - this is where uncaptured stderr will go
      out_dir=$OPTARG
      qsub_opts+=" -e $out_dir"
      ;;
  p)  # project name
      project=$OPTARG
      qsub_opts+=" -P $project"
      ;;
  r)  # cluster ram request (will be applied *per slot* if n_slots is set
      ram_req=$OPTARG
      qsub_opts+=" -l mem_free=$ram_req"
      ;;
  s)  # number of slots requested per task
```

```
# constant global options
opts_constant="-v $env_vars -cwd -r n -b n -S /bin/bash -w n -j y -q all.q "
qsub_opts+=$opts_constant


##  The command line options are supplied by Meraculous, but you can control if and h
submitter

project=
name=
out_dir=
n_tasks=
n_slots=
ram_req=
wclk_req=

while getopts e:n:o:p:r:s:t:w: opt; do
  case $opt in
  n)  # name for the job
      name=$OPTARG
      qsub_opts+=" -N $name"
      ;;
  o)  # default stdout directory - this is where uncaptured stdout plus any cluster s
      out_dir=$OPTARG
      qsub_opts+=" -o $out_dir"
      ;;
  e)  # default stderr directory - this is where uncaptured stderr will go
      out_dir=$OPTARG
      qsub_opts+=" -e $out_dir"
      ;;
  p)  # project name
      project=$OPTARG
      qsub_opts+=" -P $project"
      ;;
  r)  # cluster ram request (will be applied *per slot* if n_slots is set
      ram_req=$OPTARG
      qsub_opts+=" -l mem_free=$ram_req"
      ;;
  s)  # number of slots requested per task
"Meraculous/Current_build/bin/cluster_submit.sh" 109L, 2596C written
```

# Coverage issues

```
 2015/05/05 23:32:52 M_Utility.pm M_Utility::run_local_cmd 726> Local command:  /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/memtime -f /campusdata/BME235/S15_assemblies/
Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_mercount/mercount.hist.memtime cat /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s/
/meraculous_mercount/hist.tmp/mercount.0.*.hist | grep "|" | perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/unique.pl - 2 7 | sort -n -k 1 | awk '{print $1,$3}' > /cam
pusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_mercount/mercount.hist 2> sys_cmd.20150505-233252.err

 2015/05/05 23:32:55 meraculous.pl main::run_single_cmd 3558> memtime: (/afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/memtime -f /campusdata/BME235/S15_assemblies/Meraculo
us/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_mercount/mercount.hist.memtime cat /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meracul
ous_mercount/hist.tmp/mercount.0.*.hist | grep "|" | perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/unique.pl - 2 7 | sort -n -k 1 | awk '{print $1,$3}' > /campusdata/
BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_mercount/mercount.hist 2> sys_cmd.20150505-233252.err) 0.00 user, 0.06 system, 3.01 elapsed -- Max VS
ize = 100940KB, Max RSS = 620KB

 2015/05/05 23:32:55 M_Utility.pm M_Utility::run_local_cmd 726> Local command:  echo "set terminal png; set output '/campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-0
5-04_18h40m47s//meraculous_mercount/mercount.png'; set log y; set xlabel 'k-mer freq'; set ylabel 'nr. of distinct k-mers'; plot [2:1000] '/campusdata/BME235/S15_assemblies/Meraculous/Kmer_
61_with_more_mem_2015-05-04_18h40m47s//meraculous_mercount/mercount.hist' using 1:2 with boxes" | gnuplot 2> sys_cmd.20150505-233255.err

 2015/05/05 23:32:56 M_Utility.pm M_Utility::run_local_cmd 726> Local command:  cat /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_merco
unt/mercount.hist | awk '$1 > 2' | perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/kmerHistAnalyzer.pl -S -h -  > /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_w
ith_more_mem_2015-05-04_18h40m47s//meraculous_mercount/kha.plot 2> sys_cmd.20150505-233256.err

 2015/05/05 23:32:56 M_Utility.pm M_Utility::run_local_cmd 726> Local command:  echo "set terminal png; set output 'kha.png'; set log x; set xlabel 'k-mer depth / peak depth'; set ylabel '%
 k-mers'; plot 'kha.plot' using 1:2" | gnuplot 2> sys_cmd.20150505-233256.err

 2015/05/05 23:32:57 meraculous.pl main::check_outputs_stage_meraculous_mercount 1440> Stage validation failure:  Insufficient peak k-mer depth - must be at least 15x


 2015/05/05 23:32:57 meraculous.pl main::check_outputs_stage_meraculous_mercount 1441>
Total 61-mers (over 2x) : 53655061612
Total unique sequeces (over 2x) : 3807083376
Weighted average 61-mer depth : 14.0934821523068
```

```
    open( H, "<$assembly_dir/meraculous_mercount/mercount.hist" );

    my $volume = 0; # nr of kmers
    my $cntU = 0;  # nr of *unique* sequences of size k
    while (<H>)
    {
        my @line = split(/\s+/, $_);
        next unless ($line[0] > $meraculous_min_depth_cutoff);
        $cntU += $line[1];
        $volume += ( $line[0]*$line[1] );
    }
    close H;
    my $merDepthWAvg = $volume / $cntU;

    my $stageValidFile = "$assembly_dir/meraculous_mercount/stage_validation.$date";
    open( F, ">$stageValidFile" );

    print F "Total ${mer_size}-mers (over ${meraculous_min_depth_cutoff}x) : $volume\n";
    print F "Total unique sequeces (over ${meraculous_min_depth_cutoff}x) : $cntU\n";
    print F "Weighted average ${mer_size}-mer depth : $merDepthWAvg\n \n";
    close F;
    unless( $merDepthWAvg > 15) {    # require that we have at least 15x of mer depth
        $pLog->error( "Stage validation failure:  Insufficient peak k-mer depth - must be at least 15x \n");
        $pLog->error( "\n", `cat $stageValidFile`);
        return JGI_FAILURE;
    }

    return JGI_SUCCESS;
}
```

```
    open( H, "<$assembly_dir/meraculous_mercount/mercount.hist" );

    my $volume = 0; # nr of kmers
    my $cntU = 0;  # nr of *unique* sequences of size k
    while (<H>)
    {
        my @line = split(/\s+/, $_);
        next unless ($line[0] > $meraculous_min_depth_cutoff);
        $cntU += $line[1];
        $volume += ( $line[0]*$line[1] );
    }
    close H;
    my $merDepthWAvg = $volume / $cntU;

    my $stageValidFile = "$assembly_dir/meraculous_mercount/stage_validation.$date";
    open( F, ">$stageValidFile" );

    print F "Total ${mer_size}-mers (over ${meraculous_min_depth_cutoff}x) : $volume\n";
    print F "Total unique sequeces (over ${meraculous_min_depth_cutoff}x) : $cntU\n";
    print F "Weighted average ${mer_size}-mer depth : $merDepthWAvg\n \n";
    close F;
    unless( $merDepthWAvg > 0) {    # require that we have at least 15x of mer depth
        $pLog->error( "Stage validation failure:  Insufficient peak k-mer depth - must be at least 15x \n");
        $pLog->error( "\n", `cat $stageValidFile`);
        return JGI_FAILURE;
    }

    return JGI_SUCCESS;
}
```

# Contigs

```
[chkcole@campusrocks2 Meraculous]$ cd Kmer_61_with_more_mem_2015-05-04_18h40m47s/
[chkcole@campusrocks2 Kmer_61_with_more_mem_2015-05-04_18h40m47s]$ cd
checkpoints/         meraculous_bubble/   meraculous_import/   meraculous_mergraph/ run-kill             run_pids.history     run-term
log/                 meraculous_contigs/  meraculous_mercount/ meraculous_ufx/      run-pause            run-resume
[chkcole@campusrocks2 Kmer_61_with_more_mem_2015-05-04_18h40m47s]$ cd meraculous_contigs/
[chkcole@campusrocks2 meraculous_contigs]$ ls -lt
total 9291440
drwxrwxr-x 3 chkcole bme235           6 May  6 23:12 JOB_SET_DIR.UUtigger
-rw-rw-r-- 1 chkcole bme235           0 May  6 23:12 CLEANED-UP.1
-rw-r--r-- 1 chkcole bme235 11511131590 May  6 23:11 UUtigs.err
-rw-r--r-- 1 chkcole bme235  6346112085 May  6 23:07 UUtigs.fa.cea
-rw-r--r-- 1 chkcole bme235  4159690597 May  6 23:06 UUtigs.fa
-rw-rw-r-- 1 chkcole bme235         533 May  6 19:56 linkedScript.template.submit.20150506-185635.err
-rw-rw-r-- 1 chkcole bme235         259 May  6 19:56 linkedScript.template
-rw-rw-r-- 1 chkcole bme235           0 May  6 19:56 sys_cmd.20150506-185635.err
[chkcole@campusrocks2 meraculous_contigs]$ grep -c ">" UUtigs.fa
28610138
[chkcole@campusrocks2 meraculous_contigs]$ ls -lth
total 8.9G
drwxrwxr-x 3 chkcole bme235    6 May  6 23:12 JOB_SET_DIR.UUtigger
-rw-rw-r-- 1 chkcole bme235    0 May  6 23:12 CLEANED-UP.1
-rw-r--r-- 1 chkcole bme235  11G May  6 23:11 UUtigs.err
-rw-r--r-- 1 chkcole bme235 6.0G May  6 23:07 UUtigs.fa.cea
-rw-r--r-- 1 chkcole bme235 3.9G May  6 23:06 UUtigs.fa
-rw-rw-r-- 1 chkcole bme235  533 May  6 19:56 linkedScript.template.submit.20150506-185635.err
-rw-rw-r-- 1 chkcole bme235  259 May  6 19:56 linkedScript.template
-rw-rw-r-- 1 chkcole bme235    0 May  6 19:56 sys_cmd.20150506-185635.err
[chkcole@campusrocks2 meraculous_contigs]$
```

# Bubble-popping

```
2015/05/09 16:40:09 meraculous.pl main::run_single_cmd 3547> command passed to run_local_cmd : /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/memtime -f /campusdata/BME235
/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa.memtime perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/bubbleFin
der.pl -c /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtigs.fa.cea -f /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_w
ith_more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtigs.fa -d /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/UUtigs.mer_dept
h > /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa 2> /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_mor
e_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.err

2015/05/09 16:40:33 M_Utility.pm M_Utility::run_local_cmd 726> Local command:  /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/memtime -f /campusdata/BME235/S15_assemblies/
Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa.memtime perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/bubbleFinder.pl -c /campu
sdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtigs.fa.cea -f /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_201
5-05-04_18h40m47s//meraculous_contigs/UUtigs.fa -d /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/UUtigs.mer_depth > /campusdata/
BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa 2> /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04
_18h40m47s//meraculous_bubble/bubbletigs.err

2015/05/09 16:40:33 M_Utility.pm M_Utility::run_local_cmd 770> Command failed! Return value: (-1) Command: (/afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/memtime -f /camp
usdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa.memtime perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/
bin/bubbleFinder.pl -c /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtigs.fa.cea -f /campusdata/BME235/S15_assemblies/Meracul
ous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtigs.fa -d /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/UU
tigs.mer_depth > /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa 2> /campusdata/BME235/S15_assemblies/Meraculous/Kme
r_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.err)
For more clues on what went wrong please examine the output and stderr files produced by this command!

2015/05/09 16:40:33 meraculous.pl main::run_single_cmd 3553> ERROR: System command at line 1964 failed:         (Called from run_meraculous_bubble)

2015/05/09 16:40:33 meraculous.pl main:: 769> Stage meraculous_bubble failed (12666.956266 seconds)
2 at /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/meraculous.pl line 3553.


2015/05/09 16:40:33 meraculous.pl main:: 824> ERRORS ENCOUNTERED!

2015/05/09 16:40:33 meraculous.pl main:: 828> Total run time: 12672.162845 seconds.
[chkcole@campusrocks2 log]$
```

```
    $$command_output_ref = `$command`;

    #
    # If the command failed, generate an error.
    #
    if ($$return_value_ref = $?)
    {
        $message = "Command failed! Return value: " .
            "($$return_value_ref) Command: ($command)\n" .
            "For more clues on what went wrong please examine the output and stderr files produced by this command!";
        $log->error($message);
        return JGI_FAILURE;
    }

    #
    # Remove any leading or trailing whitespace from the result.
    #
    std_entry_processing($command_output_ref);
}
else
{
    #
    # The function was called without a variable to store the
    # command results, so just verify that the command ran
    # successfully.
    #
    if ($$return_value_ref = system("$command"))
    {
        $message = "Command failed! Return value: " .
            "($$return_value_ref, \$\! $!) Command: ($command)\n" .
            "For more clues on what went wrong please examine the output and stderr files produced by this command!";
        $log->error($message);
        return JGI_FAILURE;
    }
}
```

```
 2015/05/10 00:31:18 M_Utility.pm M_Utility::run_local_cmd 726> Local command:  /afs/cats.ucsc.edu/users/q/chkc
ole/Meraculous/Current_build/bin/memtime -f /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_
2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa.memtime perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculou
s/Current_build/bin/bubbleFinder.pl -c /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-
05-04_18h40m47s//meraculous_contigs/UUtigs.fa.cea -f /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_
more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtigs.fa -d /campusdata/BME235/S15_assemblies/Meraculous/Kme
r_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/UUtigs.mer_depth > /campusdata/BME235/S15_assemblies
/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa 2> /campusdata/BME235/S
15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.err

 2015/05/10 00:31:18 M_Utility.pm M_Utility::run_local_cmd 770> Command failed! Return value: (-1, $! Cannot al
locate memory) Command: (/afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/memtime -f /campusdata
/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bubble/bubbletigs.fa.m
emtime perl /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/bubbleFinder.pl -c /campusdata/BME2
35/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtigs.fa.cea -f /c
ampusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_contigs/UUtig
s.fa -d /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s//meraculous_bub
ble/UUtigs.mer_depth > /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-04_18h40m47s/
/meraculous_bubble/bubbletigs.fa 2> /campusdata/BME235/S15_assemblies/Meraculous/Kmer_61_with_more_mem_2015-05-
04_18h40m47s//meraculous_bubble/bubbletigs.err)
For more clues on what went wrong please examine the output and stderr files produced by this command!

 2015/05/10 00:31:18 meraculous.pl main::run_single_cmd 3553> ERROR: System command at line 1964 failed:
(Called from run_meraculous_bubble)

 2015/05/10 00:31:18 meraculous.pl main:: 769> Stage meraculous_bubble failed (16573.33281 seconds)
2 at /afs/cats.ucsc.edu/users/q/chkcole/Meraculous/Current_build/bin/meraculous.pl line 3553.


 2015/05/10 00:31:18 meraculous.pl main:: 824> ERRORS ENCOUNTERED!

 2015/05/10 00:31:18 meraculous.pl main:: 828> Total run time: 16575.497392 seconds.
[chkcole@campusrocks2 log]$
```

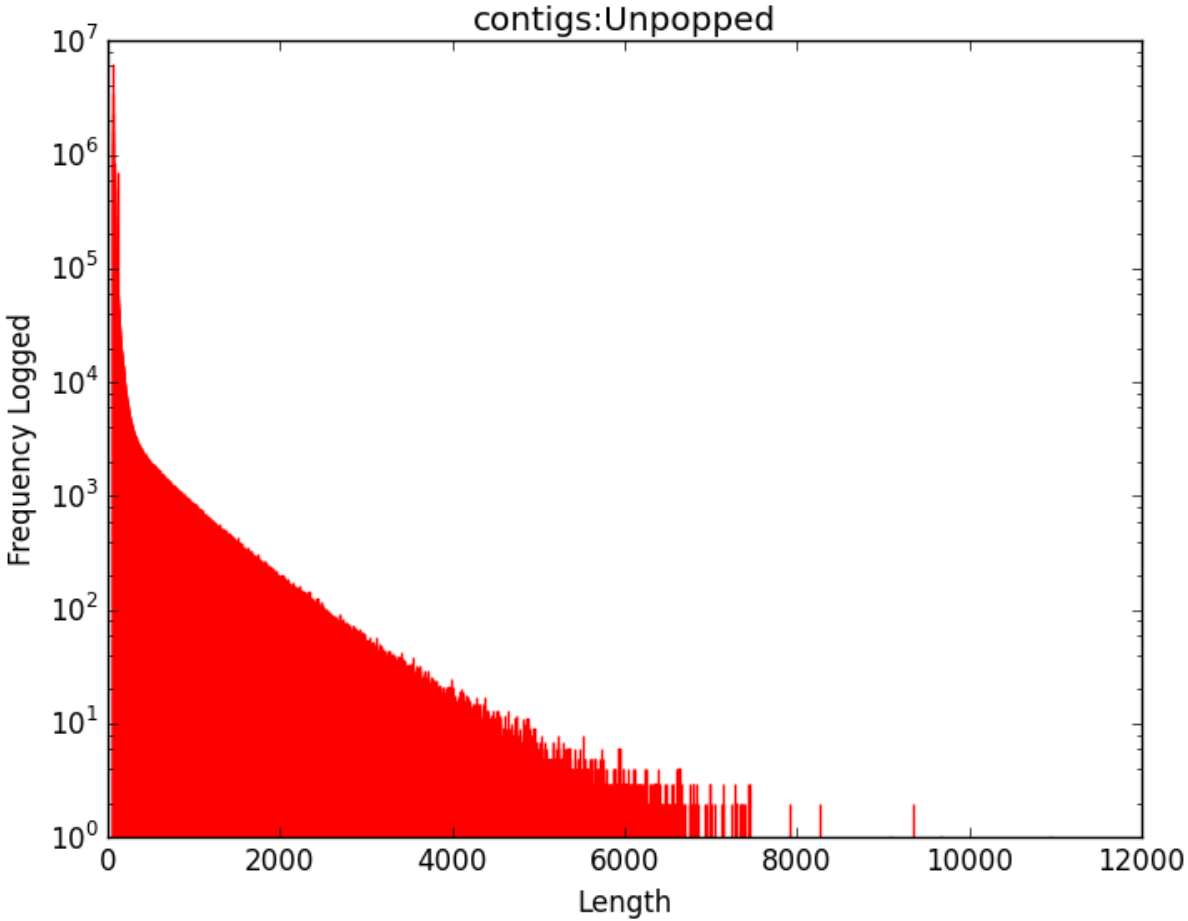
# Results so far

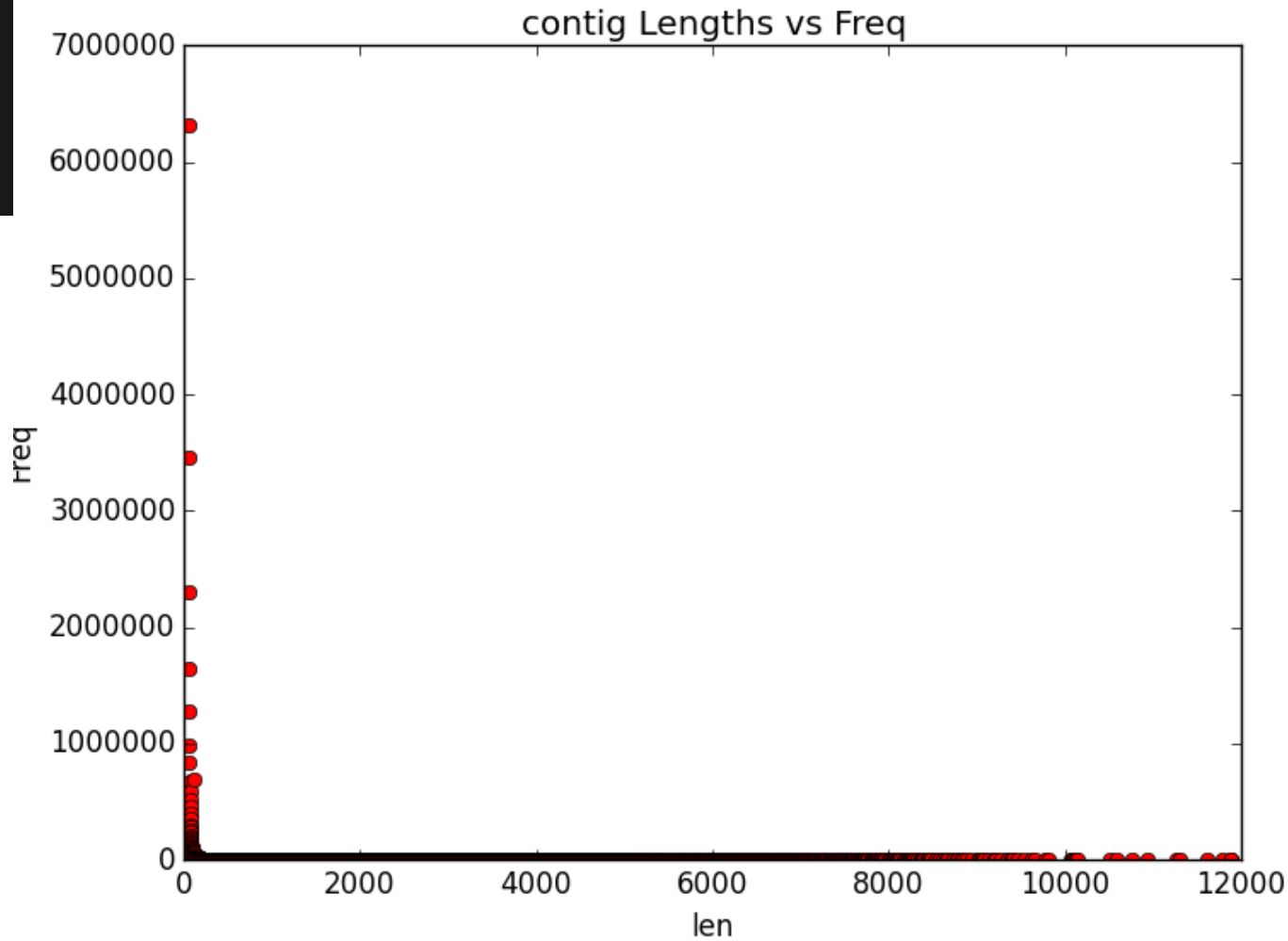
Summary statistics of contigs before bubble popping:

28,610,138 total contigs

542,137 (1.89%) contigs > 1000bp

15 (5.2e-5%) contigs > 10,000bp


contigs:Unpopped
Frequency Logged
$10^7$
$10^6$
$10^5$
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
0
2000
4000
6000
8000
10000
12000
Length


contig Lengths vs Freq
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Freq
0
2000
4000
6000
8000
10000
12000
len

# Error Correction- Quake

- Should have results soon, program is running, lots of packages to install and requires jellyfish to run correctly. Will post kmergenie results to see if musket error correction is better/worse.
- Like musket, classifies kmers as trusted or untrusted, but counts “q-mers” rather than kmers which are just kmers with quality scores magically weighted in.
- Three steps: q-mer counting (with jellyfish), cut-off calculation and finally error correction.

# Error Correction - BLESS

- BLoom-filter-based Error correction Solution for high-throughput Sequencing reads
- Uses a single minimum-sized Bloom filter (a space-efficient probabilistic data structure)
- Still compiling
  - New version came out yesterday
  - Current problem:

```
“correct_errors.cpp:949:62: error: invalid conversion from ‘const void*’ to ‘void*’ [-fpermissive]
    MPI_File_write_at(f_out, write_offset, mmap.data(), residue, MPI_CHAR, &status)”
```

# Error Correction - Racer

- Rapid and Accurate Correction of Errors in Reads
- Uses k-mer counts
  - k-mers with counts above threshold are deemed correct
  - Different approach than Musket, BLESS and Quake, which use k-mer spectrum
- Not available online
  - Requested from Dr. Lucian Ilie from the University of Western Ontario
  - Location:  /campusdata/BME235/bin/racer

# Preqc of new data

- Currently running

# Next Steps

- Finish Meraculous' bubble popping step
- Assess with CEGMA
- Re-run
  - New (error corrected) data
  - Incorporate scaffold from new mate-pair data
- Improve assembly with GapCloser and REAPR
- Meta-assembly
- Annotate genome
- Publish
- Rock banana slug t-shirts