Digital Normalization

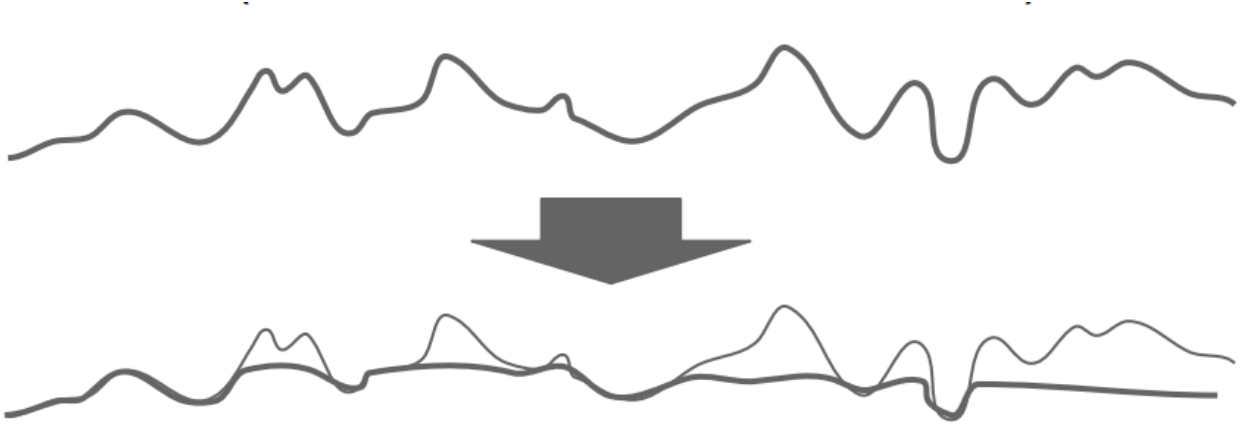
16 Oct 15

Recap

- Form Hypothesis
- Collect Seq Data
- QC
- Error Correct
- Trimming
- Normalize
- QC
- Assemble
  - Genome v. transcriptome
- QC
- Post-assembly
  - BLAST/HMM
- Biology

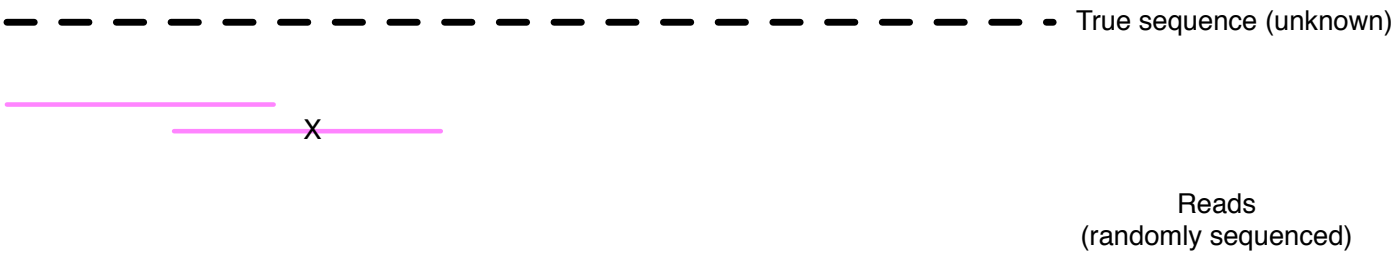
Digital Normalization

Perfect Storm of data analysis – What to do???

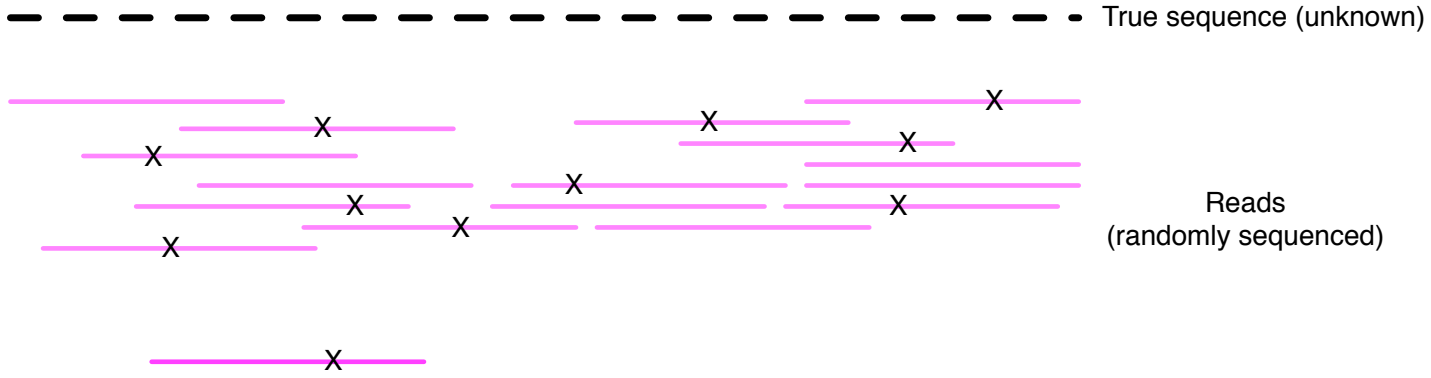


Brown 2012 arXiv:1203.4802v2

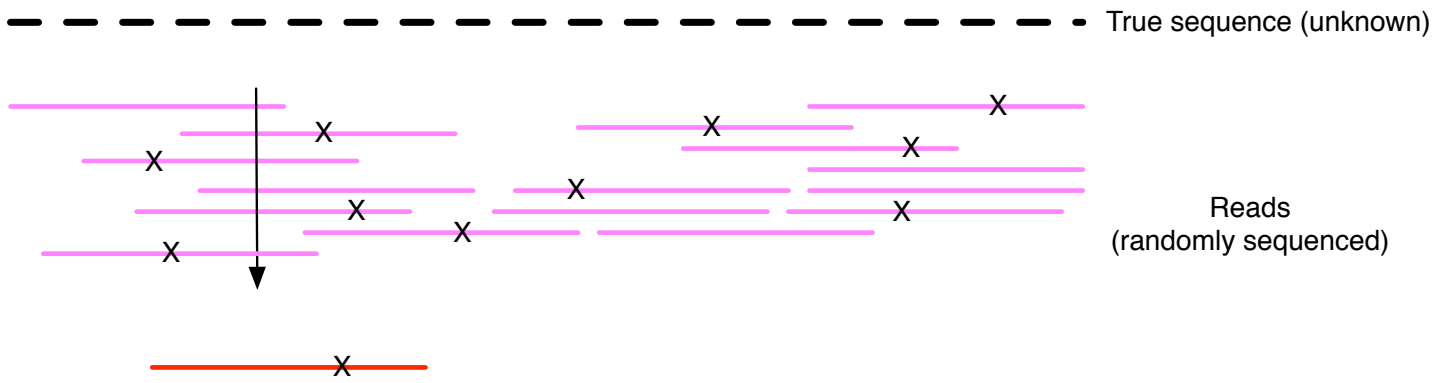
Digital Normalization



Digital Normalization

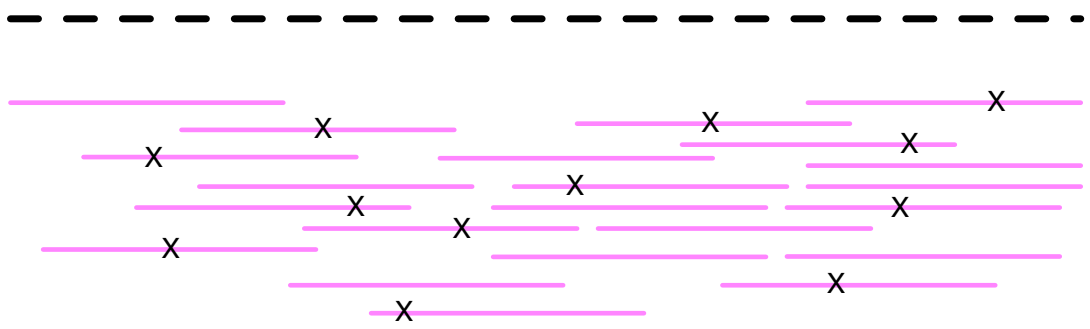


Digital Normalization



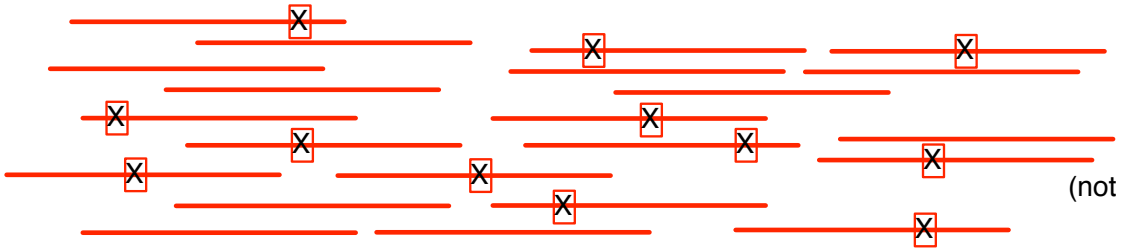

```
for read in dataset:
    if estimated_coverage(read) < C:
        accept(read)
    else:
        discard(read)
```

Digital Normalization



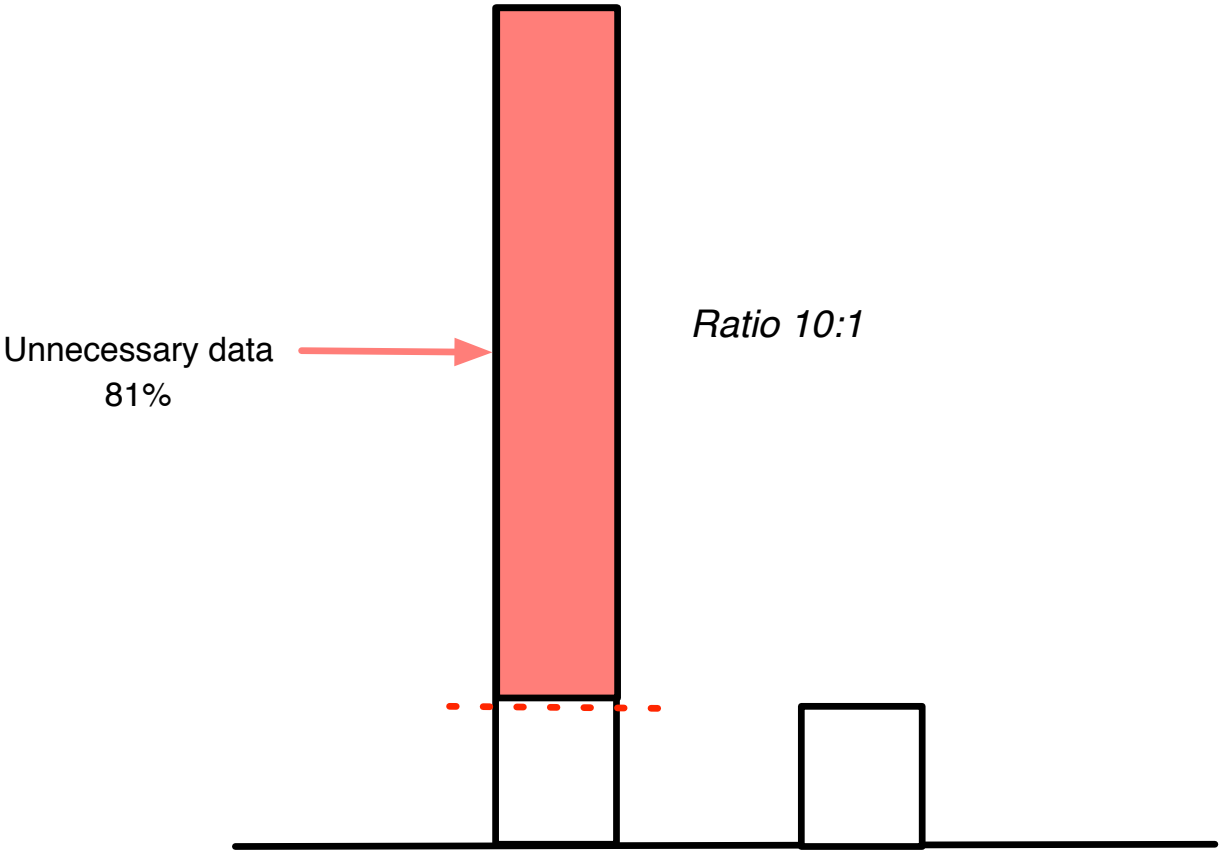
Reads
(randomly sequenced)

```
for read in dataset:
   if estimated_coverage(read) < C:
      accept(read)
   else:
      discard(read)
```



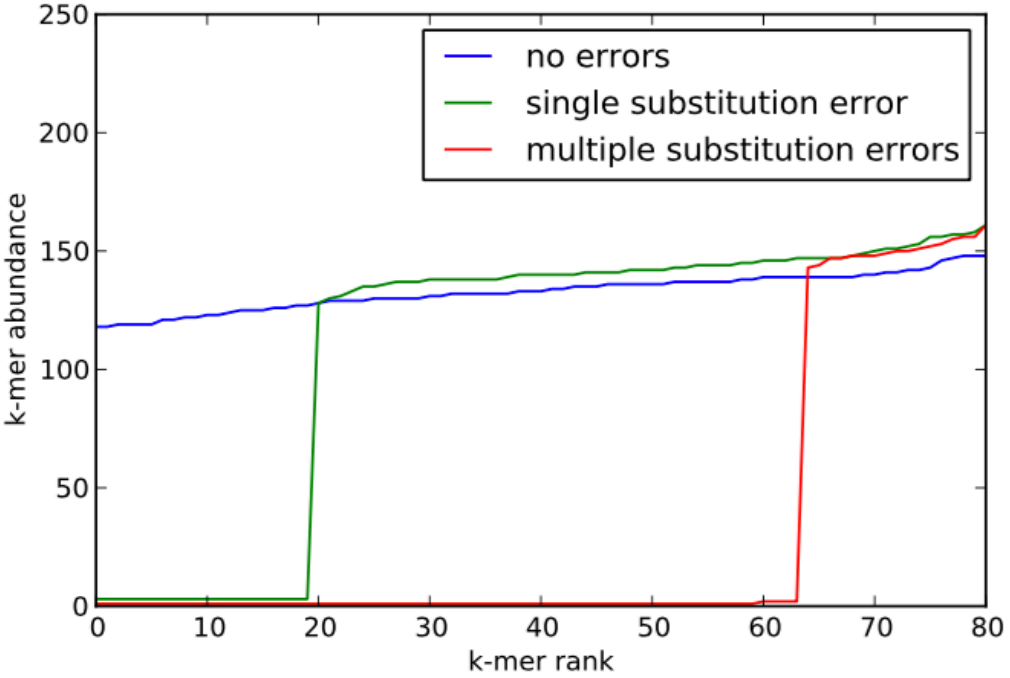
Redundant reads
(not needed for assembly)

Digital Normalization



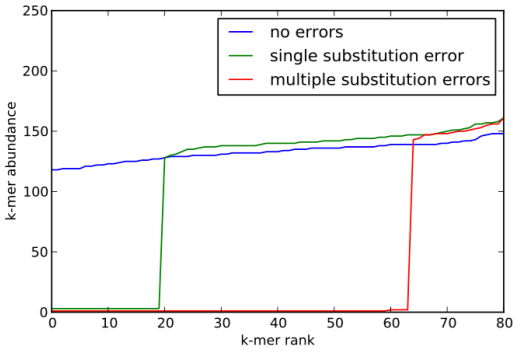
Digital Normalization

```
for read in dataset:
   if estimated_coverage(read) < C:
      accept(read)
   else:
      discard(read)
```



Digital Normalization

No error



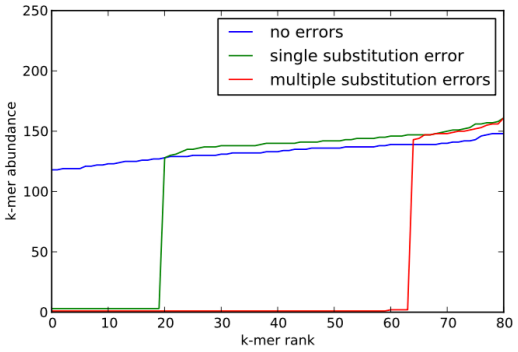
3mer freq.

CAT=32
ATG=34
TGC=36
GCA=35
CAT=33
ATT=34
TTG=40

CATGCATTG
CAT
ATG
TGC
GCA
CAT
ATT
TTG

Digital Normalization

1error



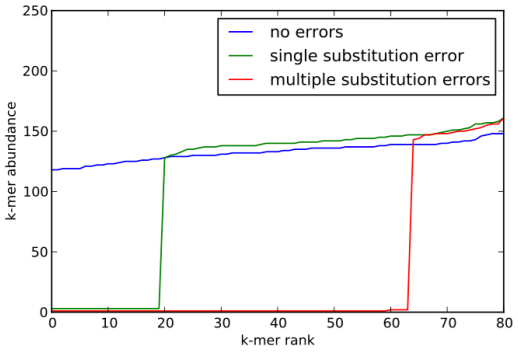
3mer freq.

CAT=32
ATG=34
TG**A**=1
G**A**A=1
AAT=1
ATT=34
TTG=40

CATG**A**ATTG
CAT
ATG
TG**A**
G**A**A
AAT
ATT
TTG

Digital Normalization

>1 error

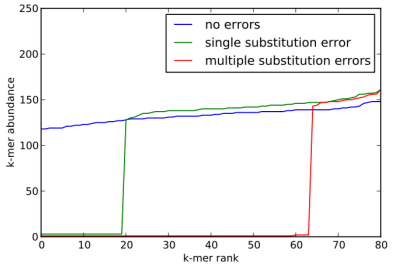


3mer freq.

CAT=32
ATG=34
TG**A**=1
G**A**A=1
AAT=1
AT**C**=1
T**C**G=1

CATG**A**AT**C**G
CAT
ATG
TG**A**
G**A**A
AAT
AT**C**
T**C**G

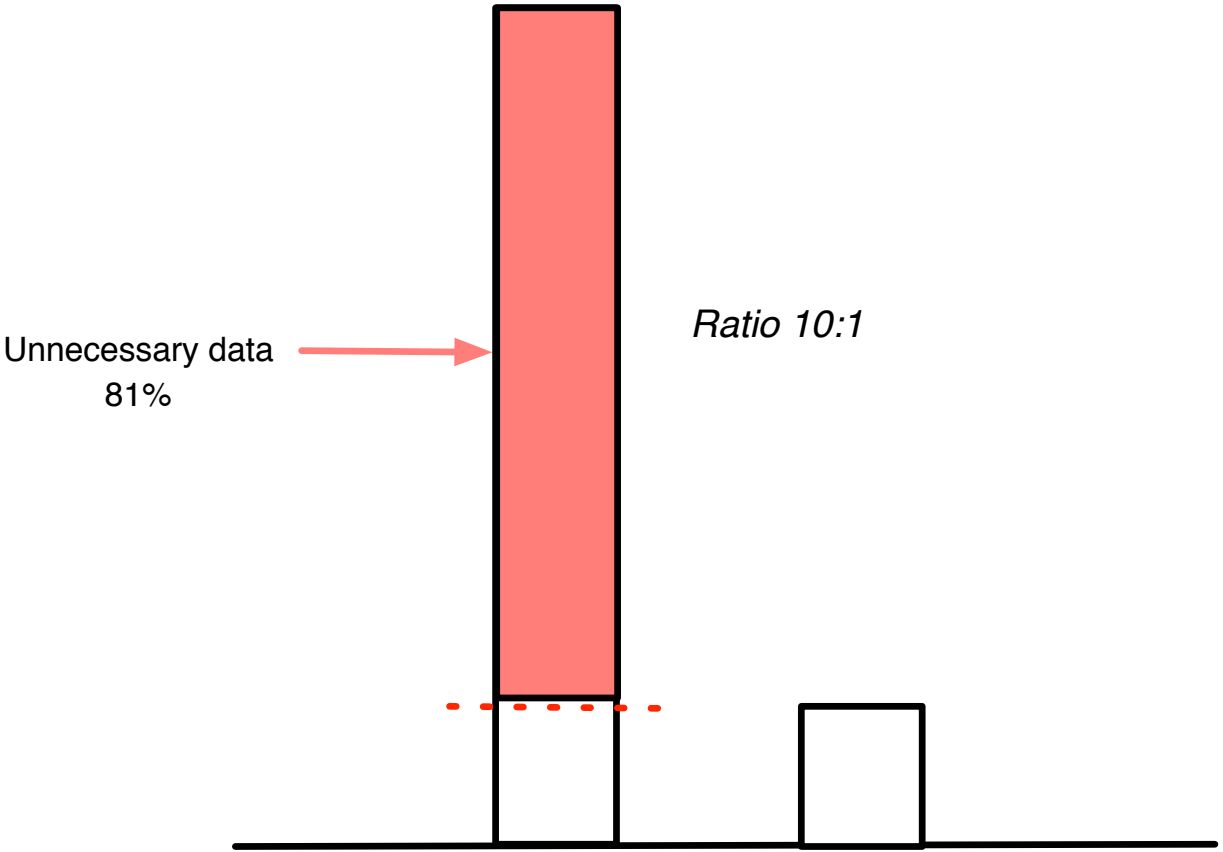
Digital Normalization



Median kmer abundance

0 error: 32,33,34,34,35,36,40
1 error: 1,1,1,32,24,24,40
>1 error: 1,1,1,1,1,32,34

Digital Normalization



Digital Normalization

Table 1. Digital normalization to C=20 removes many erroneous k-mers from sequencing data sets. Numbers in parentheses indicate number of true k-mers lost at each step, based on reference.

Data set	True 20-mers	20-mers in reads	20-mers at C=20	% reads kept
Simulated genome	399,981	8,162,813	3,052,007 (-2)	19%
Simulated mRNAseq	48,100	2,466,638 (-88)	1,087,916 (-9)	4.1%
E. coli genome	4,542,150	175,627,381 (-152)	90,844,428 (-5)	11%
Yeast mRNAseq	10,631,882	224,847,659 (-683)	10,625,416 (-6,469)	9.3%
Mouse mRNAseq	43,830,642	709,662,624 (-23,196)	43,820,319 (-13,400)	26.4%

Digital Normalization

Table 3. Three-pass digital normalization reduces computational requirements for contig assembly of genomic data.

Data set	N reads pre/post	Assembly time pre/post	Assembly memory pre/post
E. coli	31m / 0.6m	1040s / 63s (16.5x)	11.2gb / 0.5 gb (22.4x)
S. aureus single-cell	58m / 0.3m	5352s / 35s (153x)	54.4gb / 0.4gb (136x)
Deltaproteobacteria single-cell	67m / 0.4m	4749s / 26s (182.7x)	52.7gb / 0.4gb (131.8x)

Table 4. Single-pass digital normalization to C=20 reduces computational requirements for transcriptome assembly.

Data set	N reads pre/post	Assembly time pre/post	Assembly memory pre/post
Yeast (Oases)	100m / 9.3m	181 min / 12 min (15.1x)	45.2gb / 8.9gb (5.1x)
Yeast (Trinity)	100m / 9.3m	887 min / 145 min (6.1x)	31.8gb / 10.4gb (3.1x)
Mouse (Oases)	100m / 26.4m	761 min/ 73 min (10.4x)	116.0gb / 34.6gb (3.4x)
Mouse (Trinity)	100m / 26.4m	2297 min / 634 min (3.6x)	42.1gb / 36.4gb (1.2x)