Metagenomics

Ravin Poudel

02 July, 2019

QC run for metagenomics analyses

```
# used libraries
library(tidyverse)
```

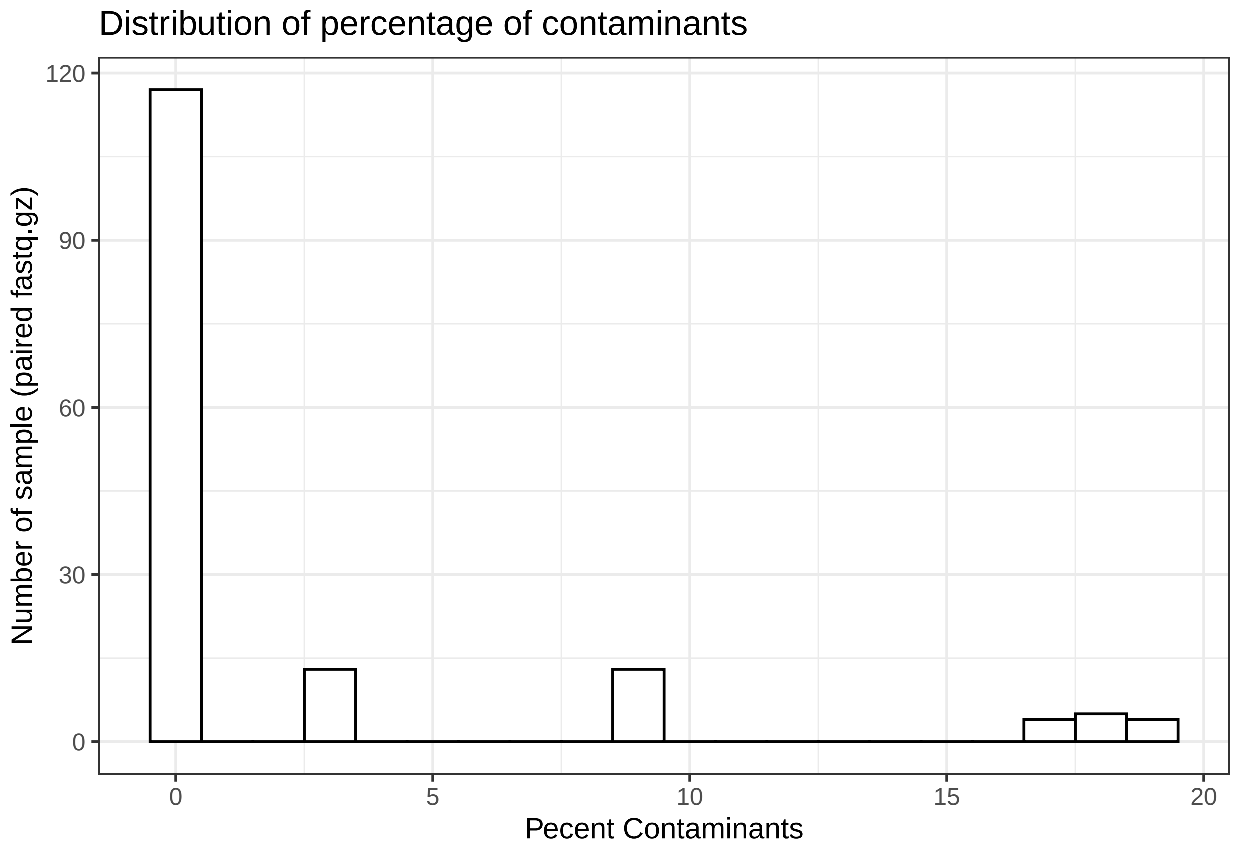
```
## -- Attaching packages ----------------------------------------------------------------------------- ti

## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------- tic
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

```
# read in data
data = read.csv("parse_json_md.csv", header=T, row.names = 1)
head(data)
```

```
##         SampleID    SampleName Treatment GeoLocation TotalReads TotalBases
## 0  A1_run2_lane3 A1_run2_lane3        A1        AFRS    4541932  679870331
## 4  A1_run3_lane4 A1_run3_lane4        A1        AFRS    4292178  644365708
## 10 A1_run1_lane1 A1_run1_lane1        A1        AFRS     102848   15472289
## 18 A1_run4_lane3 A1_run4_lane3        A1        AFRS    4657772  699116858
## 52 A1_run3_lane2 A1_run3_lane2        A1        AFRS    4260714  639681080
## 65 A1_run3_lane3 A1_run3_lane3        A1        AFRS    4403216  661034461
##    Contaminants Percent_Contaminants
## 0            21              0.00046
## 4            21              0.00049
## 10            0              0.00000
## 18            8              0.00017
## 52           10              0.00023
## 65           17              0.00039
```

```
# Distribution of contaminants
ggplot(data, aes(x=Percent_Contaminants)) +
  geom_histogram(color="black", fill="white", binwidth = 1) +
  theme_bw() +
  labs(title="Distribution of percentage of contaminants",
       y ="Number of sample (paired fastq.gz)", x = "Pecent Contaminants")
```

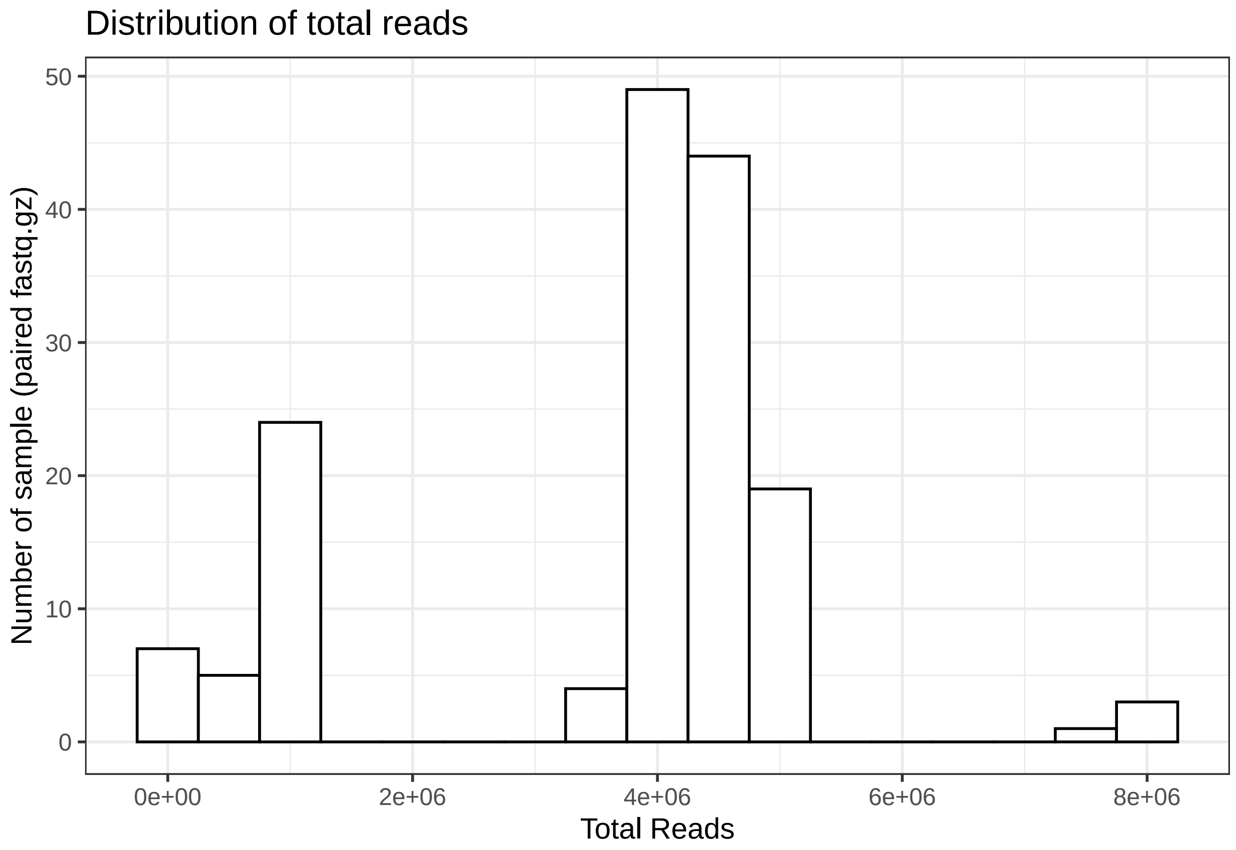


```r
df <- data %>%
  summarise(mean_PC = mean(Percent_Contaminants),
            sd_PC = sd(Percent_Contaminants),
            median_PC = median(Percent_Contaminants),
            min_PC = min(Percent_Contaminants),
            max_PC = max(Percent_Contaminants))

kable(df,caption = "Percentage contamination for overall reads", format = "pandoc", align = 'c')
```

Table 1: Percentage contamination for overall reads

mean_PC	sd_PC	median_PC	min_PC	max_PC
2.51288	5.34847	0.000635	0	18.82453

```r
# Distribution of total reads
ggplot(data, aes(x=TotalReads)) +
  geom_histogram(color="black", fill="white", binwidth = 500000) +
  theme_bw() +
  labs(title="Distribution of total reads",
       y ="Number of sample (paired fastq.gz)", x = "Total Reads")
```

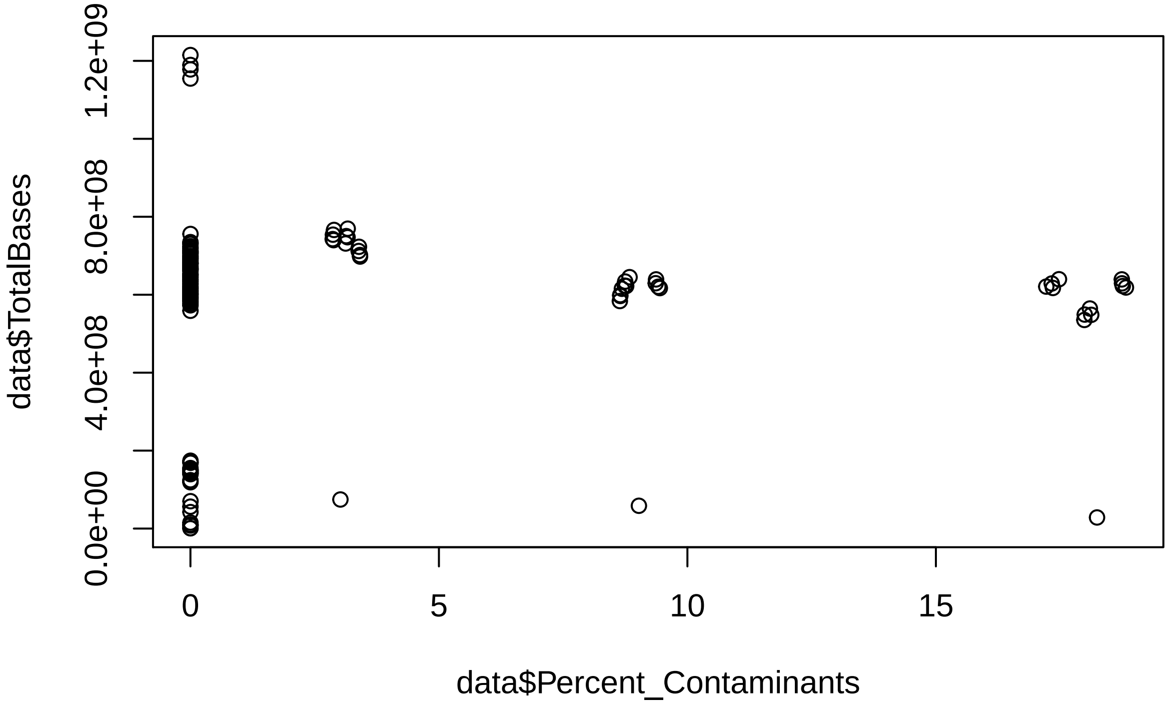


```r
df2 <- data %>%
  summarise(mean_TR = mean(TotalReads),
            sd_TR = sd(TotalReads),
            median_TR = median(TotalReads),
            min_TR = min(TotalReads),
            max_TR = max(TotalReads),
            total_TR = sum(TotalReads))

kable(df2,caption = "Summary of total reads", format = "pandoc", align = 'c')
```

Table 2: Summary of total reads

mean_TR	sd_TR	median_TR	min_TR	max_TR	total_TR
3599739	1717995	4148077	4924	8117650	561559256

```r
plot(data$Percent_Contaminants,data$TotalBases)
```



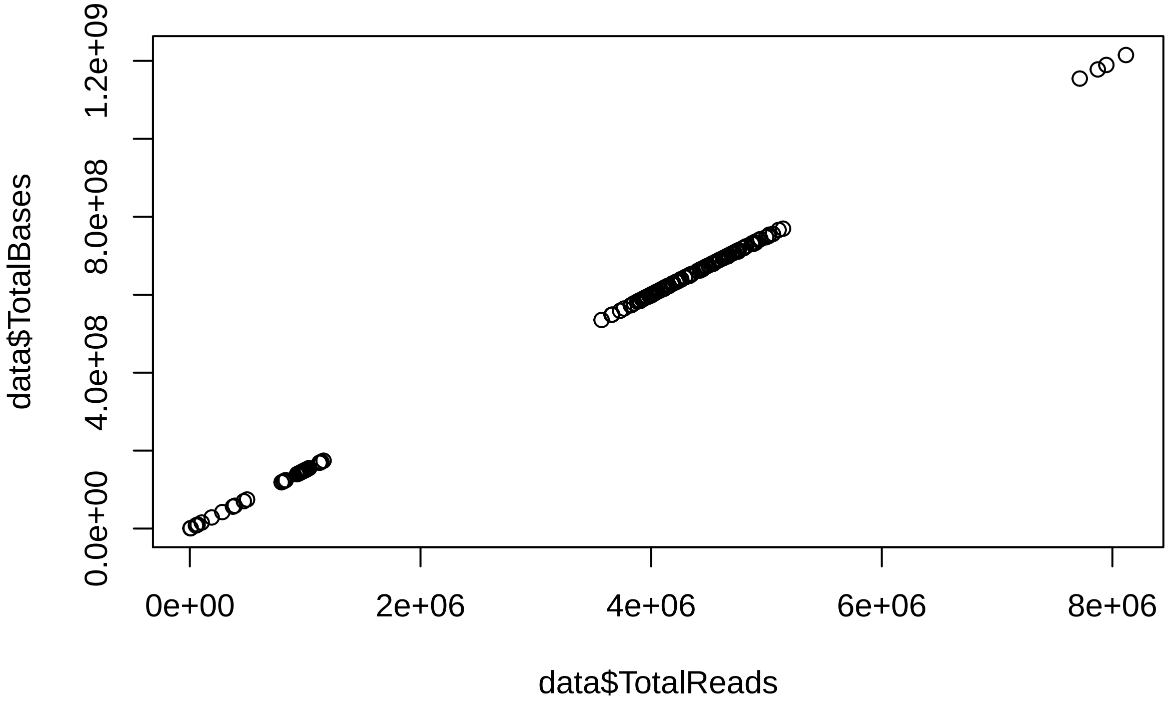
```
lin_PC = lm(data$Percent_Contaminants ~ data$TotalBases)
summary(lin_PC)
```

```
##
## Call:
## lm(formula = data$Percent_Contaminants ~ data$TotalBases)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.437 -2.681 -2.577 -1.316 16.429
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.774e+00  9.984e-01   1.777   0.0776 .
## data$TotalBases 1.369e-09  1.670e-09   0.820   0.4137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.354 on 154 degrees of freedom
## Multiple R-squared:  0.004343,   Adjusted R-squared:  -0.002122
## F-statistic: 0.6718 on 1 and 154 DF,  p-value: 0.4137
```

```
cor(data$Percent_Contaminants,data$TotalBases)
```

```
## [1] 0.06590242
```

```
plot(data$TotalReads, data$TotalBases)
```



```
linm = lm(data$TotalReads ~ data$TotalBases)
summary(linm)
```

```
##
## Call:
## lm(formula = data$TotalReads ~ data$TotalBases)
##
## Residuals:
##    Min      1Q Median     3Q    Max
##  -6553  -4732  -1552   2373  16049
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -2.518e+03  1.108e+03   -2.272   0.0245 *
## data$TotalBases  6.673e-03  1.854e-06 3598.866   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5943 on 154 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.295e+07 on 1 and 154 DF,  p-value: < 2.2e-16
```

```
cor(data$TotalReads, data$TotalBases)
```

```
## [1] 0.9999941
```

Grouped by location/ treatments

```
data_byTreatment <- data %>%
  group_by(Treatment) %>%
  summarise(mean_TR = mean(TotalReads),
            sd_TR = sd(TotalReads),
            mean_TB = mean(TotalBases),
            sd_TB = sd(TotalBases),
            mean_PC = mean(Percent_Contaminants),
            sd_PC = sd(Percent_Contaminants),
            nSample = n())


kable(data_byTreatment,caption = "Summary based on grouped_by_treatment", format = "pandoc", align = 'c
```

Table 3: Summary based on grouped_by_treatment

Treatment	mean_TR	sd_TR	mean_TB	sd_TB	mean_PC	sd_PC	nSample
A1	4106484.8	1208531.5	615836905	181228573	0.0003146	0.0001314	13
A2	5007806.3	2310272.2	750639685	345190139	0.0004569	0.0002190	13
A3	3705160.2	1098138.3	555713485	164675800	0.0005554	0.0008642	13
C1W	4572616.6	1234445.5	685471041	184925463	3.1312038	0.2176653	13
C2W	3711686.3	1087213.0	557059438	163235232	18.0695092	0.5865727	13
C3W	3839497.4	1042472.8	575666992	156347148	8.9484969	0.3280179	13
M1	4032462.3	1197446.2	605205297	179736070	0.0001000	0.0000819	13
M2	867818.3	273964.7	130070792	41154937	0.0003115	0.0003074	13
M3	940469.4	294267.7	141128767	44177522	0.0011123	0.0006063	13
O1S	3655517.8	991679.0	548328442	148747044	0.0005669	0.0003307	13
O2S	4356581.1	1229231.0	653406189	184344142	0.0005869	0.0001827	13
O3S	4400765.4	1195679.5	659455379	179001964	0.0013431	0.0003287	13

Summary based on individual fastq.

```
data_bySample <- data %>%
  select(SampleName, Treatment, GeoLocation,TotalReads, TotalBases, Percent_Contaminants) %>%
  arrange(Treatment)


kable(data_bySample,caption = "Summary based on individual fastq", format = "pandoc", align = 'c')
```

Table 4: Summary based on individual fastq

SampleName	Treatment	GeoLocation	TotalReads	TotalBases	Percent_Contaminants
A1_run2_lane3	A1	AFRS	4541932	679870331	0.00046
A1_run3_lane4	A1	AFRS	4292178	644365708	0.00049
A1_run1_lane1	A1	AFRS	102848	15472289	0.00000
A1_run4_lane3	A1	AFRS	4657772	699116858	0.00017
A1_run3_lane2	A1	AFRS	4260714	639681080	0.00023
A1_run3_lane3	A1	AFRS	4403216	661034461	0.00039
A1_run2_lane4	A1	AFRS	4448052	665799283	0.00029
A1_run4_lane4	A1	AFRS	4524830	679301308	0.00031
A1_run4_lane2	A1	AFRS	4498124	675176538	0.00031
A1_run2_lane2	A1	AFRS	4333160	648411889	0.00032
A1_run2_lane1	A1	AFRS	4424766	662145823	0.00034

SampleName	Treatment	GeoLocation	TotalReads	TotalBases	Percent_Contaminants
A1_run3_lane1	A1	AFRS	4334060	650640629	0.00032
A1_run4_lane1	A1	AFRS	4562650	684863574	0.00046
A2_run4_lane4	A2	AFRS	4219082	633482940	0.00062
A2_run2_lane1	A2	AFRS	7873516	1177981523	0.00029
A2_run4_lane2	A2	AFRS	4184056	628099712	0.00081
A2_run3_lane4	A2	AFRS	4078526	612404261	0.00051
A2_run3_lane3	A2	AFRS	4178394	627388551	0.00055
A2_run2_lane3	A2	AFRS	8117650	1214885117	0.00023
A2_run3_lane1	A2	AFRS	4100592	615697496	0.00046
A2_run4_lane3	A2	AFRS	4340152	651514073	0.00053
A2_run3_lane2	A2	AFRS	4044528	607326835	0.00064
A2_run2_lane4	A2	AFRS	7947828	1189441888	0.00043
A2_run4_lane1	A2	AFRS	4247460	637606761	0.00064
A2_run2_lane2	A2	AFRS	7717088	1154567193	0.00023
A2_run1_lane1	A2	AFRS	52610	7919555	0.00000
A3_run3_lane1	A3	AFRS	3949156	592764297	0.00033
A3_run4_lane2	A3	AFRS	3965400	595101407	0.00045
A3_run4_lane3	A3	AFRS	4109590	616713312	0.00022
A3_run1_lane1	A3	AFRS	58614	8820132	0.00341
A3_run2_lane1	A3	AFRS	4035322	604277392	0.00022
A3_run4_lane1	A3	AFRS	4024862	604007529	0.00037
A3_run4_lane4	A3	AFRS	4004098	601038428	0.00025
A3_run3_lane2	A3	AFRS	3894862	584672123	0.00023
A3_run3_lane4	A3	AFRS	3920610	588505986	0.00031
A3_run3_lane3	A3	AFRS	4011408	602127307	0.00057
A3_run2_lane3	A3	AFRS	4154864	622346259	0.00024
A3_run2_lane4	A3	AFRS	4076546	610638622	0.00039
A3_run2_lane2	A3	AFRS	3961750	593262505	0.00023
C1W_run2_lane3	C1W	Israel1	5143998	769479521	3.16476
C1W_run1_lane1	C1W	Israel1	496320	74620526	3.01842
C1W_run3_lane2	C1W	Israel1	4650188	697917263	3.41350
C1W_run4_lane4	C1W	Israel1	4947834	742670147	2.86299
C1W_run3_lane4	C1W	Israel1	4676800	701886399	3.41163
C1W_run4_lane1	C1W	Israel1	5024802	754086585	2.86803
C1W_run3_lane3	C1W	Israel1	4817368	722976674	3.39183
C1W_run2_lane2	C1W	Israel1	4890746	731380124	3.12300
C1W_run4_lane3	C1W	Israel1	5105580	766200123	2.88915
C1W_run3_lane1	C1W	Israel1	4744680	712057165	3.38562
C1W_run2_lane4	C1W	Israel1	4997062	747489626	3.16184
C1W_run2_lane1	C1W	Israel1	5018000	750388866	3.13448
C1W_run4_lane2	C1W	Israel1	4930638	739970517	2.88040
C2W_run2_lane3	C2W	Israel1	3765202	564619987	18.09871
C2W_run2_lane1	C2W	Israel1	3662480	549038878	17.99546
C2W_run4_lane1	C2W	Israel1	4188184	628791812	17.33104
C2W_run2_lane4	C2W	Israel1	3657090	548399118	18.12570
C2W_run1_lane1	C2W	Israel1	189242	28481892	18.24225
C2W_run3_lane1	C2W	Israel1	4190544	629201920	18.74485
C2W_run3_lane3	C2W	Israel1	4257894	639331842	18.74112
C2W_run3_lane4	C2W	Israel1	4144486	622306551	18.76030
C2W_run2_lane2	C2W	Israel1	3570492	535236834	17.98769
C2W_run4_lane3	C2W	Israel1	4261136	639746355	17.47679
C2W_run3_lane2	C2W	Israel1	4119220	618522220	18.82453

SampleName	Treatment	GeoLocation	TotalReads	TotalBases	Percent_Contaminants
C2W_run4_lane4	C2W	Israel1	4134338	620789833	17.22073
C2W_run4_lane2	C2W	Israel1	4111614	617305448	17.35445
C3W_run2_lane4	C3W	Israel1	3989880	596783121	8.65044
C3W_run2_lane2	C3W	Israel1	3904036	583774890	8.64085
C3W_run4_lane3	C3W	Israel1	4297384	645053525	8.83645
C3W_run3_lane4	C3W	Israel1	4133306	620456406	9.41370
C3W_run3_lane3	C3W	Israel1	4258864	639286679	9.36992
C3W_run3_lane1	C3W	Israel1	4193166	629420057	9.36254
C3W_run3_lane2	C3W	Israel1	4109922	616959267	9.44534
C3W_run2_lane3	C3W	Israel1	4113258	615253364	8.68144
C3W_run4_lane1	C3W	Israel1	4222092	633749448	8.74836
C3W_run4_lane2	C3W	Israel1	4145410	622244195	8.76869
C3W_run4_lane4	C3W	Israel1	4150744	623127841	8.74046
C3W_run1_lane1	C3W	Israel1	389196	58511655	9.02399
C3W_run2_lane1	C3W	Israel1	4006208	599050443	8.64828
M1_run2_lane1	M1	Miller	4225444	633464487	0.00009
M1_run3_lane1	M1	Miller	4349312	653107679	0.00000
M1_run2_lane2	M1	Miller	4141842	620916291	0.00010
M1_run4_lane2	M1	Miller	4425844	664452363	0.00009
M1_run4_lane1	M1	Miller	4479962	672567590	0.00020
M1_run2_lane3	M1	Miller	4352602	652726631	0.00014
M1_run4_lane4	M1	Miller	4470182	671232307	0.00018
M1_run3_lane2	M1	Miller	4297838	645430521	0.00014
M1_run3_lane3	M1	Miller	4429286	665123136	0.00005
M1_run3_lane4	M1	Miller	4327018	649781429	0.00005
M1_run4_lane3	M1	Miller	4597230	690152395	0.00026
M1_run2_lane4	M1	Miller	4258242	638595856	0.00000
M1_run1_lane1	M1	Miller	67208	10118178	0.00000
M2_run3_lane1	M2	Miller	992284	148933589	0.00040
M2_run1_lane1	M2	Miller	7512	1128846	0.00000
M2_run2_lane4	M2	Miller	815082	121782662	0.00037
M2_run3_lane3	M2	Miller	1009884	151580375	0.00020
M2_run2_lane2	M2	Miller	795398	118801065	0.00025
M2_run4_lane3	M2	Miller	1034874	155283935	0.00010
M2_run4_lane2	M2	Miller	999618	150001127	0.00030
M2_run3_lane2	M2	Miller	978686	146914011	0.00051
M2_run4_lane4	M2	Miller	1006366	151045245	0.00119
M2_run2_lane1	M2	Miller	810774	121107827	0.00000
M2_run4_lane1	M2	Miller	1015292	152353433	0.00039
M2_run3_lane4	M2	Miller	984960	147847133	0.00010
M2_run2_lane3	M2	Miller	830908	124141047	0.00024
M3_run3_lane2	M3	Miller	1124138	168809628	0.00240
M3_run2_lane4	M3	Miller	982002	147225872	0.00071
M3_run2_lane2	M3	Miller	950206	142413510	0.00189
M3_run4_lane2	M3	Miller	929722	139553834	0.00108
M3_run3_lane4	M3	Miller	1129566	169613737	0.00124
M3_run3_lane3	M3	Miller	1159552	174115853	0.00086
M3_run4_lane4	M3	Miller	936454	140594906	0.00085
M3_run4_lane3	M3	Miller	967764	145253997	0.00062
M3_run3_lane1	M3	Miller	1140618	171267547	0.00088
M3_run2_lane3	M3	Miller	996470	149397676	0.00110
M3_run2_lane1	M3	Miller	962682	144290372	0.00166

SampleName	Treatment	GeoLocation	TotalReads	TotalBases	Percent_Contaminants
M3_run4_lane1	M3	Miller	942004	141395861	0.00117
M3_run1_lane1	M3	Miller	4924	741183	0.00000
O1S_run3_lane4	O1S	Israel2	4000080	600492166	0.00030
O1S_run2_lane3	O1S	Israel2	3927766	588404042	0.00031
O1S_run3_lane3	O1S	Israel2	4120596	618571719	0.00053
O1S_run4_lane3	O1S	Israel2	4008002	601466016	0.00040
O1S_run2_lane2	O1S	Israel2	3731052	558762848	0.00048
O1S_run3_lane1	O1S	Israel2	4055880	608847437	0.00064
O1S_run1_lane1	O1S	Israel2	374030	56277326	0.00000
O1S_run4_lane2	O1S	Israel2	3851288	577977913	0.00117
O1S_run4_lane1	O1S	Israel2	3935724	590626400	0.00084
O1S_run2_lane4	O1S	Israel2	3823526	572779738	0.00044
O1S_run4_lane4	O1S	Israel2	3880292	582449246	0.00116
O1S_run2_lane1	O1S	Israel2	3830072	573601148	0.00050
O1S_run3_lane2	O1S	Israel2	3983424	598013750	0.00060
O2S_run3_lane3	O2S	Israel2	4896132	734987949	0.00071
O2S_run3_lane1	O2S	Israel2	4824172	724173630	0.00035
O2S_run4_lane3	O2S	Israel2	4711860	707014488	0.00076
O2S_run4_lane2	O2S	Israel2	4537590	680941085	0.00055
O2S_run4_lane4	O2S	Israel2	4583828	688022679	0.00061
O2S_run2_lane2	O2S	Israel2	4538278	679499369	0.00029
O2S_run2_lane4	O2S	Israel2	4667036	698992342	0.00051
O2S_run3_lane2	O2S	Israel2	4738692	711403454	0.00063
O2S_run2_lane3	O2S	Israel2	4803934	719471264	0.00027
O2S_run2_lane1	O2S	Israel2	4665128	698472324	0.00077
O2S_run1_lane1	O2S	Israel2	281636	42368716	0.00071
O2S_run3_lane4	O2S	Israel2	4759218	714464316	0.00080
O2S_run4_lane1	O2S	Israel2	4628050	694468844	0.00067
O3S_run3_lane1	O3S	Israel2	4796326	719827303	0.00146
O3S_run4_lane3	O3S	Israel2	4622806	693302332	0.00167
O3S_run3_lane2	O3S	Israel2	4696032	704837933	0.00177
O3S_run3_lane4	O3S	Israel2	4721786	708672296	0.00110
O3S_run2_lane4	O3S	Israel2	4883990	730097747	0.00092
O3S_run2_lane2	O3S	Israel2	4755750	710789362	0.00118
O3S_run3_lane3	O3S	Israel2	4873106	731371876	0.00111
O3S_run4_lane2	O3S	Israel2	4434778	665230225	0.00115
O3S_run4_lane1	O3S	Israel2	4541118	681111510	0.00174
O3S_run2_lane3	O3S	Israel2	5056086	755856204	0.00103
O3S_run1_lane1	O3S	Israel2	467794	70294465	0.00171
O3S_run2_lane1	O3S	Israel2	4906510	733258701	0.00096
O3S_run4_lane4	O3S	Israel2	4453868	668269977	0.00166