Advanced Molecular Detection

Southeast Region Bioinformatics

R Tidyverse

09/09/2024

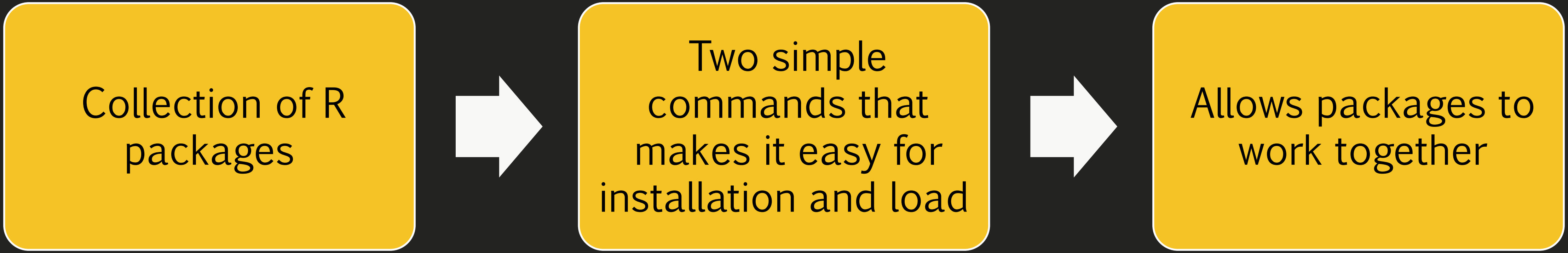
Updates

Office Hours-

- **September 16** - R Training Part 6 - ggplot2
- **September 30** - R Training Part 7 - ggtree
- **October 14** – To be determined

Advanced Molecular Detection
Southeast Region Bioinformatics

Tidyverse in a Nutshell

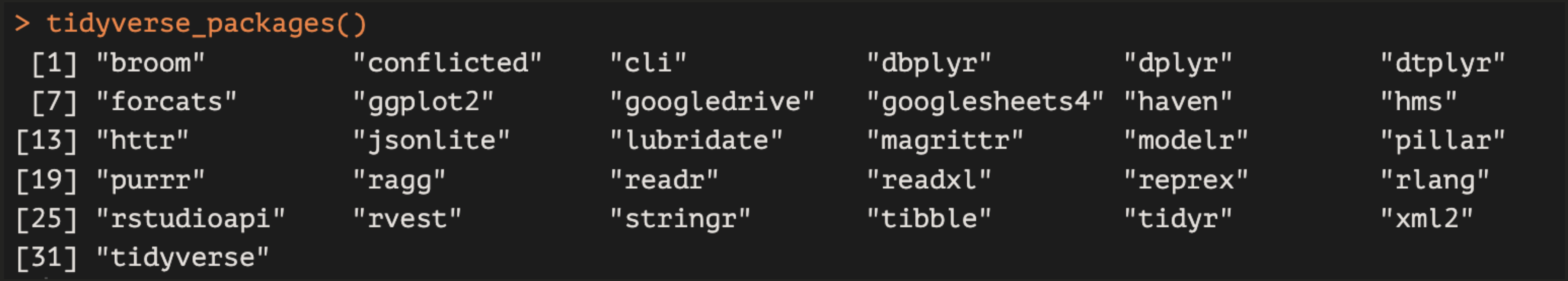


Collection of R packages

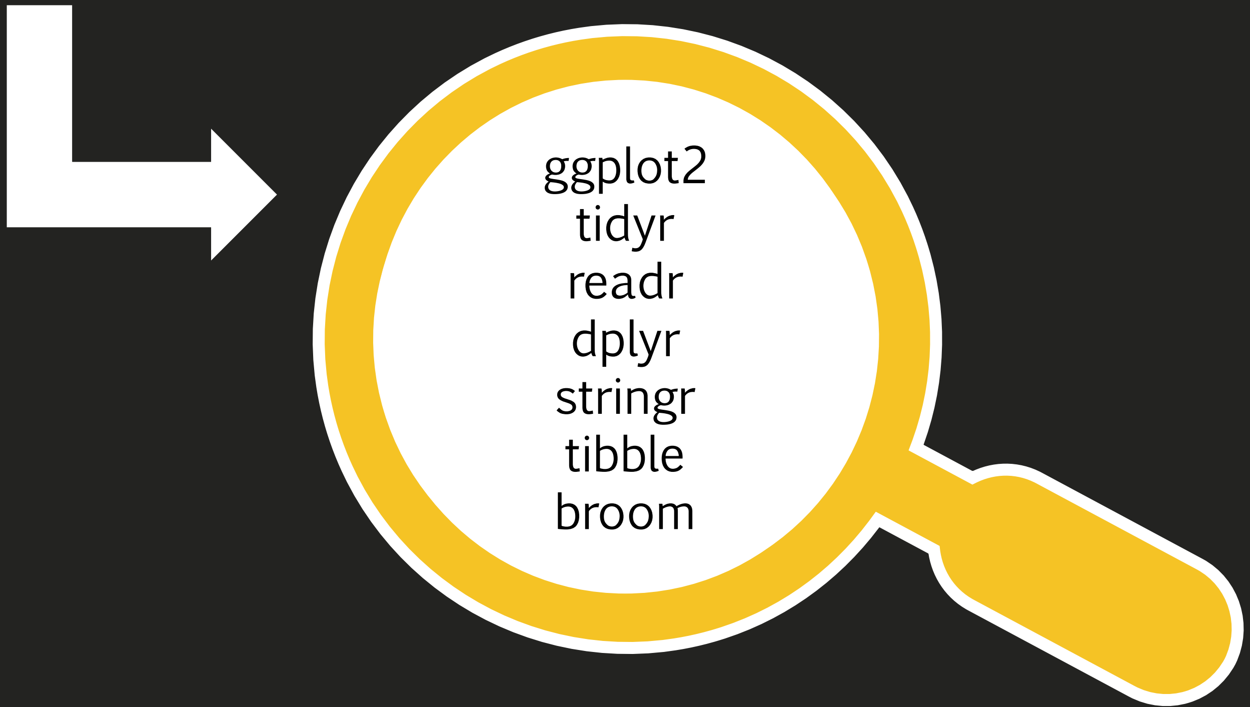
Two simple commands that makes it easy for installation and load

Allows packages to work together

Collection of Packages



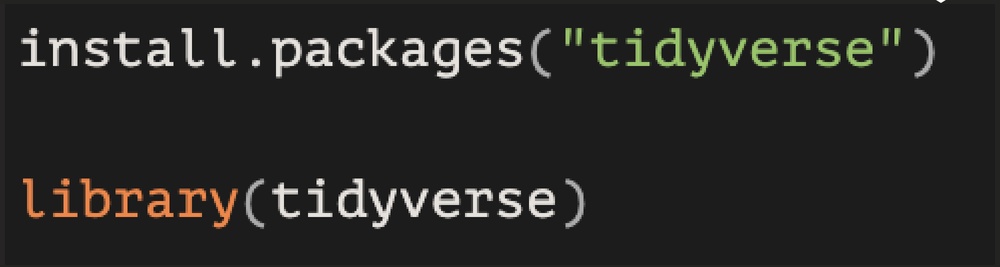
```
> tidyverse_packages()
 [1] "broom"         "conflicted"    "cli"           "dbplyr"        "dplyr"         "dtplyr"
 [7] "forcats"       "ggplot2"       "googledrive"   "googlesheets4" "haven"         "hms"
[13] "httr"          "jsonlite"      "lubridate"     "magrittr"      "modelr"        "pillar"
[19] "purrr"         "ragg"          "readr"         "readxl"        "reprex"        "rlang"
[25] "rstudioapi"    "rvest"         "stringr"       "tibble"        "tidyr"         "xml2"
[31] "tidyverse"
```



ggplot2
tidyr
readr
dplyr
stringr
tibble
broom

Two Commands

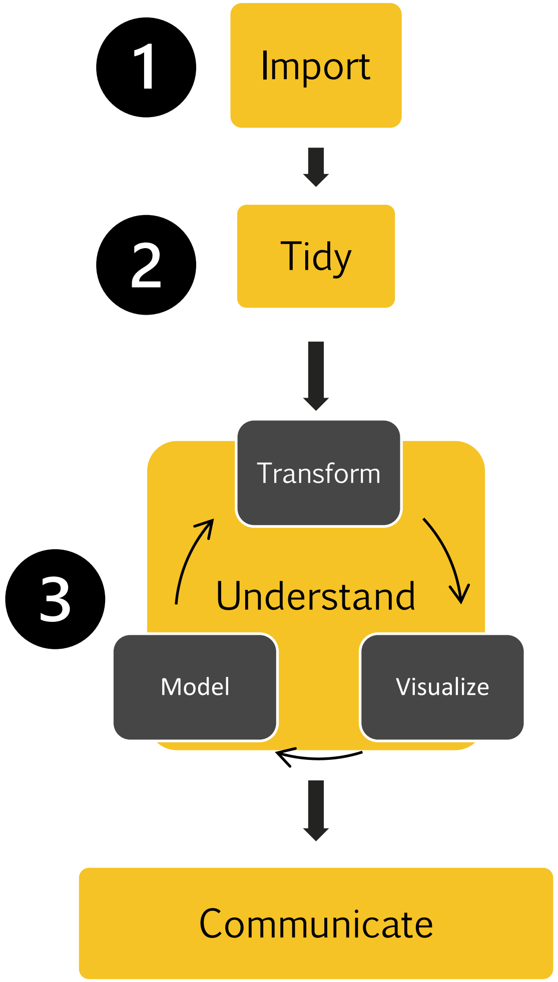
Allows us to easily install and load all those packages that was previously mentioned



```
install.packages("tidyverse")

library(tidyverse)
```

This workflow allow us to implement tidyverse, and allowing all the packages to work together

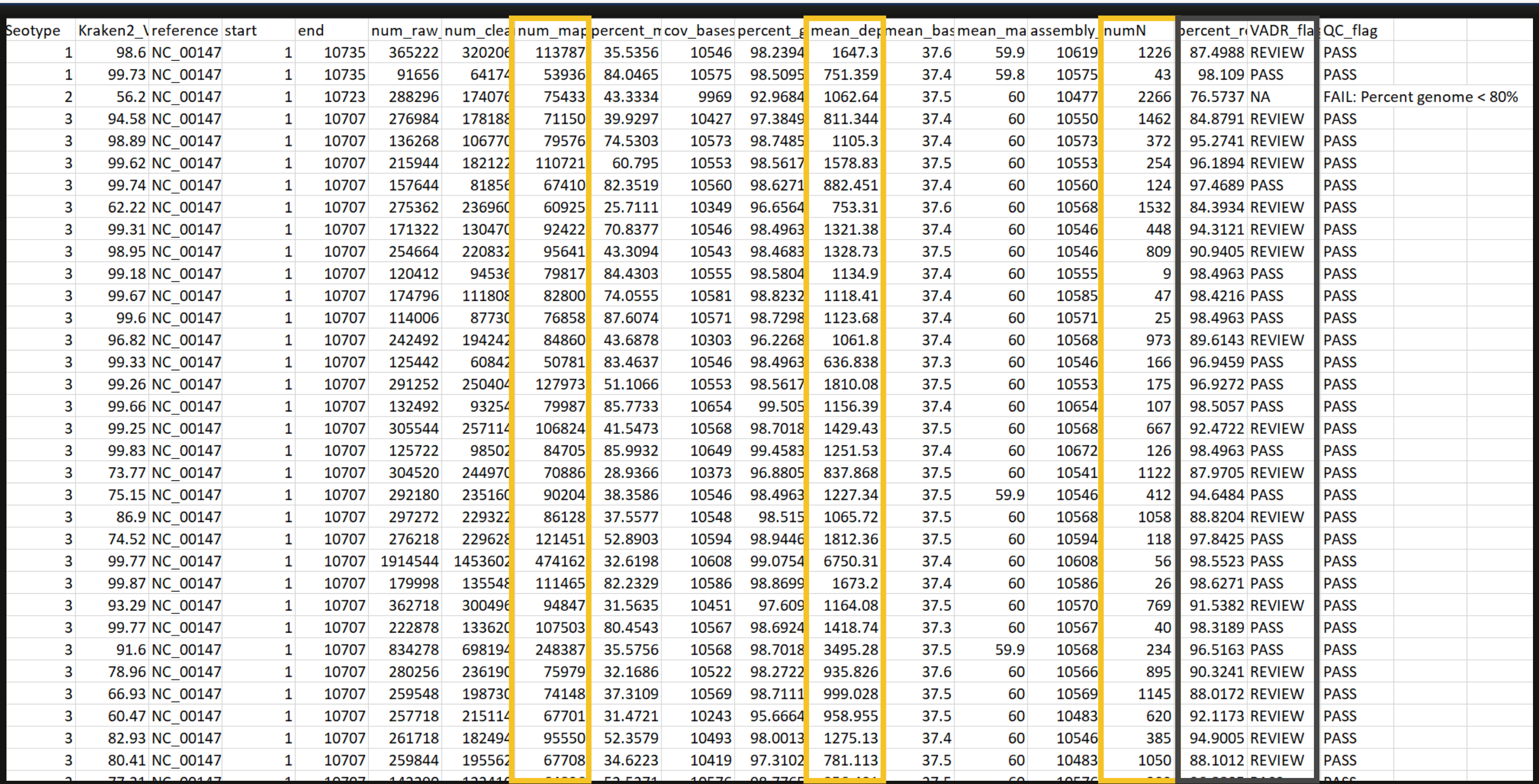
Data Analysis Workflow



Advanced Molecular Detection
Southeast Region Bioinformatics

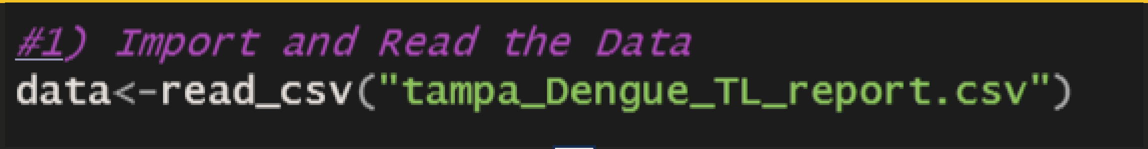
Tidyverse in Action

Example: A dataset (Tampa Dengue Report) is given, we would like to see if there is a correlation between number of Ns in the sequences to mean depth and number of mapped reads.

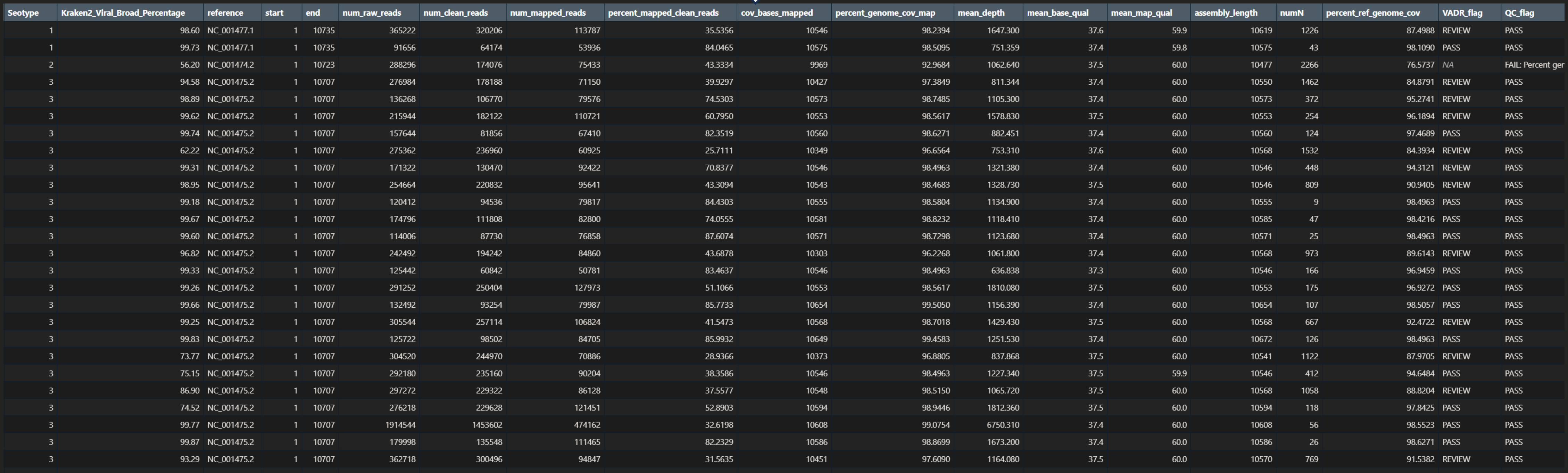


Seotype	Kraken2_\	reference	start	end	num_raw_	num_clea	num_map	percent_n	cov_bases	percent_g	mean_dep	mean_bas	mean_ma	assembly_	numN	percent_r	VADR_fla	QC_flag
1	98.6	NC_00147	1	10735	365222	320206	113787	35.5356	10546	98.2394	1647.3	37.6	59.9	10619	1226	87.4988	REVIEW	PASS
1	99.73	NC_00147	1	10735	91656	64174	53936	84.0465	10575	98.5095	751.359	37.4	59.8	10575	43	98.109	PASS	PASS
2	56.2	NC_00147	1	10723	288296	174076	75433	43.3334	9969	92.9684	1062.64	37.5	60	10477	2266	76.5737	NA	FAIL: Percent genome < 80%
3	94.58	NC_00147	1	10707	276984	178188	71150	39.9297	10427	97.3849	811.344	37.4	60	10550	1462	84.8791	REVIEW	PASS
3	98.89	NC_00147	1	10707	136268	106770	79576	74.5303	10573	98.7485	1105.3	37.4	60	10573	372	95.2741	REVIEW	PASS
3	99.62	NC_00147	1	10707	215944	182122	110721	60.795	10553	98.5617	1578.83	37.5	60	10553	254	96.1894	REVIEW	PASS
3	99.74	NC_00147	1	10707	157644	81856	67410	82.3519	10560	98.6271	882.451	37.4	60	10560	124	97.4689	PASS	PASS
3	62.22	NC_00147	1	10707	275362	236960	60925	25.7111	10349	96.6564	753.31	37.6	60	10568	1532	84.3934	REVIEW	PASS
3	99.31	NC_00147	1	10707	171322	130470	92422	70.8377	10546	98.4963	1321.38	37.4	60	10546	448	94.3121	REVIEW	PASS
3	98.95	NC_00147	1	10707	254664	220832	95641	43.3094	10543	98.4683	1328.73	37.5	60	10546	809	90.9405	REVIEW	PASS
3	99.18	NC_00147	1	10707	120412	94536	79817	84.4303	10555	98.5804	1134.9	37.4	60	10555	9	98.4963	PASS	PASS
3	99.67	NC_00147	1	10707	174796	111808	82800	74.0555	10581	98.8232	1118.41	37.4	60	10585	47	98.4216	PASS	PASS
3	99.6	NC_00147	1	10707	114006	87730	76858	87.6074	10571	98.7298	1123.68	37.4	60	10571	25	98.4963	PASS	PASS
3	96.82	NC_00147	1	10707	242492	194242	84860	43.6878	10303	96.2268	1061.8	37.4	60	10568	973	89.6143	REVIEW	PASS
3	99.33	NC_00147	1	10707	125442	60842	50781	83.4637	10546	98.4963	636.838	37.3	60	10546	166	96.9459	PASS	PASS
3	99.26	NC_00147	1	10707	291252	250404	127973	51.1066	10553	98.5617	1810.08	37.5	60	10553	175	96.9272	PASS	PASS
3	99.66	NC_00147	1	10707	132492	93254	79987	85.7733	10654	99.505	1156.39	37.4	60	10654	107	98.5057	PASS	PASS
3	99.25	NC_00147	1	10707	305544	257114	106824	41.5473	10568	98.7018	1429.43	37.5	60	10568	667	92.4722	REVIEW	PASS
3	99.83	NC_00147	1	10707	125722	98502	84705	85.9932	10649	99.4583	1251.53	37.4	60	10672	126	98.4963	PASS	PASS
3	73.77	NC_00147	1	10707	304520	244970	70886	28.9366	10373	96.8805	837.868	37.5	60	10541	1122	87.9705	REVIEW	PASS
3	75.15	NC_00147	1	10707	292180	235160	90204	38.3586	10546	98.4963	1227.34	37.5	59.9	10546	412	94.6484	PASS	PASS
3	86.9	NC_00147	1	10707	297272	229322	86128	37.5577	10548	98.515	1065.72	37.5	60	10568	1058	88.8204	REVIEW	PASS
3	74.52	NC_00147	1	10707	276218	229628	121451	52.8903	10594	98.9446	1812.36	37.5	60	10594	118	97.8425	PASS	PASS
3	99.77	NC_00147	1	10707	1914544	1453602	474162	32.6198	10608	99.0754	6750.31	37.4	60	10608	56	98.5523	PASS	PASS
3	99.87	NC_00147	1	10707	179998	135548	111465	82.2329	10586	98.8699	1673.2	37.4	60	10586	26	98.6271	PASS	PASS
3	93.29	NC_00147	1	10707	362718	300496	94847	31.5635	10451	97.609	1164.08	37.5	60	10570	769	91.5382	REVIEW	PASS
3	99.77	NC_00147	1	10707	222878	133620	107503	80.4543	10567	98.6924	1418.74	37.3	60	10567	40	98.3189	PASS	PASS
3	91.6	NC_00147	1	10707	834278	698194	248387	35.5756	10568	98.7018	3495.28	37.5	59.9	10568	234	96.5163	PASS	PASS
3	78.96	NC_00147	1	10707	280256	236190	75979	32.1686	10522	98.2722	935.826	37.6	60	10566	895	90.3241	REVIEW	PASS
3	66.93	NC_00147	1	10707	259548	198730	74148	37.3109	10569	98.7111	999.028	37.5	60	10569	1145	88.0172	REVIEW	PASS
3	60.47	NC_00147	1	10707	257718	215114	67701	31.4721	10243	95.6664	958.955	37.5	60	10483	620	92.1173	REVIEW	PASS
3	82.93	NC_00147	1	10707	261718	182494	95550	52.3579	10493	98.0013	1275.13	37.4	60	10546	385	94.9005	REVIEW	PASS
3	80.41	NC_00147	1	10707	259844	195562	67708	34.6223	10419	97.3102	781.113	37.5	60	10483	1050	88.1012	REVIEW	PASS

❶ Import



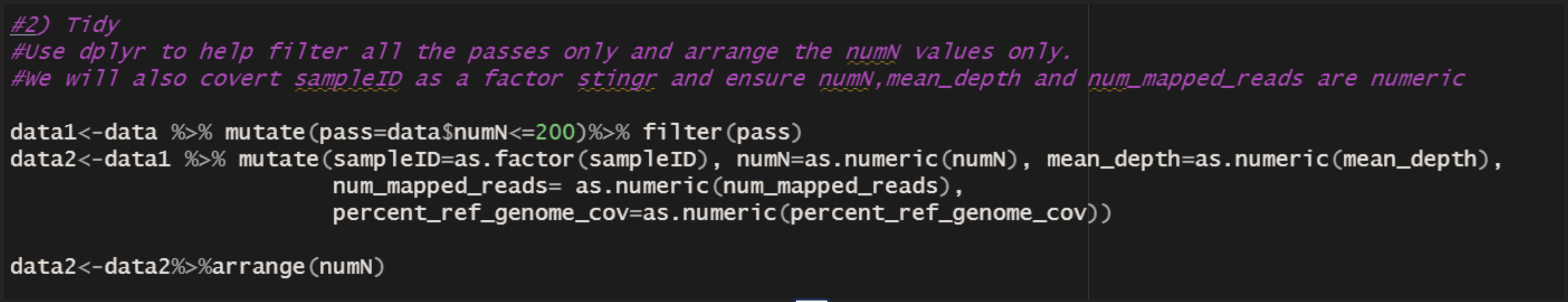
```
#1) Import and Read the Data
data<-read_csv("tampa_Dengue_TL_report.csv")
```



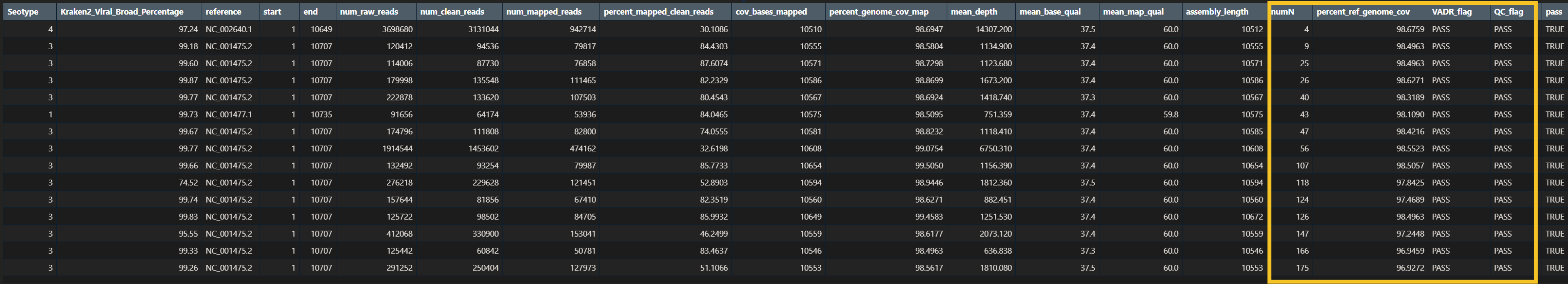
Seotype	Kraken2_Viral_Broad_Percentage	reference	start	end	num_raw_reads	num_clean_reads	num_mapped_reads	percent_mapped_clean_reads	cov_bases_mapped	percent_genome_cov_map	mean_depth	mean_base_qual	mean_map_qual	assembly_length	numN	percent_ref_genome_cov	VADR_flag	QC_flag
1	98.60	NC_001477.1	1	10735	365222	320206	113787	35.5356	10546	98.2394	1647.300	37.6	59.9	10619	1226	87.4988	REVIEW	PASS
1	99.73	NC_001477.1	1	10735	91656	64174	53936	84.0465	10575	98.5095	751.359	37.4	59.8	10575	43	98.1090	PASS	PASS
2	56.20	NC_001474.2	1	10723	288296	174076	75433	43.3334	9969	92.9684	1062.640	37.5	60.0	10477	2266	76.5737	NA	FAIL: Percent ger
3	94.58	NC_001475.2	1	10707	276984	178188	71150	39.9297	10427	97.3849	811.344	37.4	60.0	10550	1462	84.8791	REVIEW	PASS
3	98.89	NC_001475.2	1	10707	136268	106770	79576	74.5303	10573	98.7485	1105.300	37.4	60.0	10573	372	95.2741	REVIEW	PASS
3	99.62	NC_001475.2	1	10707	215944	182122	110721	60.7950	10553	98.5617	1578.830	37.5	60.0	10553	254	96.1894	REVIEW	PASS
3	99.74	NC_001475.2	1	10707	157644	81856	67410	82.3519	10560	98.6271	882.451	37.4	60.0	10560	124	97.4689	PASS	PASS
3	62.22	NC_001475.2	1	10707	275362	236960	60925	25.7111	10349	96.6564	753.310	37.6	60.0	10568	1532	84.3934	REVIEW	PASS
3	99.31	NC_001475.2	1	10707	171322	130470	92422	70.8377	10546	98.4963	1321.380	37.4	60.0	10546	448	94.3121	REVIEW	PASS
3	98.95	NC_001475.2	1	10707	254664	220832	95641	43.3094	10543	98.4683	1328.730	37.5	60.0	10546	809	90.9405	REVIEW	PASS
3	99.18	NC_001475.2	1	10707	120412	94536	79817	84.4303	10555	98.5804	1134.900	37.4	60.0	10555	9	98.4963	PASS	PASS
3	99.67	NC_001475.2	1	10707	174796	111808	82800	74.0555	10581	98.8232	1118.410	37.4	60.0	10585	47	98.4216	PASS	PASS
3	99.60	NC_001475.2	1	10707	114006	87730	76858	87.6074	10571	98.7298	1123.680	37.4	60.0	10571	25	98.4963	PASS	PASS
3	96.82	NC_001475.2	1	10707	242492	194242	84860	43.6878	10303	96.2268	1061.800	37.4	60.0	10568	973	89.6143	REVIEW	PASS
3	99.33	NC_001475.2	1	10707	125442	60842	50781	83.4637	10546	98.4963	636.838	37.3	60.0	10546	166	96.9459	PASS	PASS
3	99.26	NC_001475.2	1	10707	291252	250404	127973	51.1066	10553	98.5617	1810.080	37.5	60.0	10553	175	96.9272	PASS	PASS
3	99.66	NC_001475.2	1	10707	132492	93254	79987	85.7733	10654	99.5050	1156.390	37.4	60.0	10654	107	98.5057	PASS	PASS
3	99.25	NC_001475.2	1	10707	305544	257114	106824	41.5473	10568	98.7018	1429.430	37.5	60.0	10568	667	92.4722	REVIEW	PASS
3	99.83	NC_001475.2	1	10707	125722	98502	84705	85.9932	10649	99.4583	1251.530	37.4	60.0	10672	126	98.4963	PASS	PASS
3	73.77	NC_001475.2	1	10707	304520	244970	70886	28.9366	10373	96.8805	837.868	37.5	60.0	10541	1122	87.9705	REVIEW	PASS
3	75.15	NC_001475.2	1	10707	292180	235160	90204	38.3586	10546	98.4963	1227.340	37.5	59.9	10546	412	94.6484	PASS	PASS
3	86.90	NC_001475.2	1	10707	297272	229322	86128	37.5577	10548	98.5150	1065.720	37.5	60.0	10568	1058	88.8204	REVIEW	PASS
3	74.52	NC_001475.2	1	10707	276218	229628	121451	52.8903	10594	98.9446	1812.360	37.5	60.0	10594	118	97.8425	PASS	PASS
3	99.77	NC_001475.2	1	10707	1914544	1453602	474162	32.6198	10608	99.0754	6750.310	37.4	60.0	10608	56	98.5523	PASS	PASS
3	99.87	NC_001475.2	1	10707	179998	135548	111465	82.2329	10586	98.8699	1673.200	37.4	60.0	10586	26	98.6271	PASS	PASS
3	93.29	NC_001475.2	1	10707	362718	300496	94847	31.5635	10451	97.6090	1164.080	37.5	60.0	10570	769	91.5382	REVIEW	PASS

Advanced Molecular Detection
Southeast Region Bioinformatics

2 Tidy



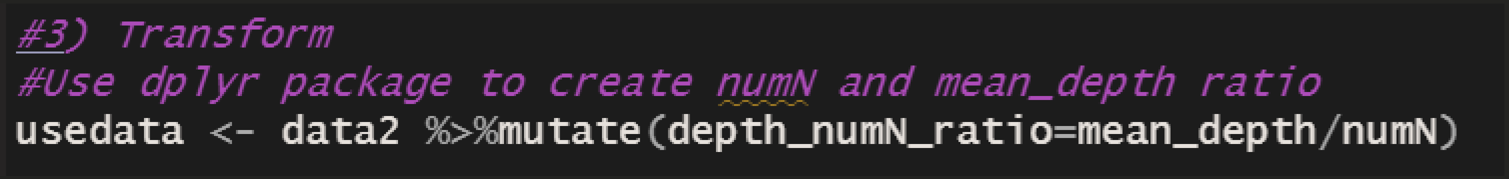
```
#2) Tidy
#Use dplyr to help filter all the passes only and arrange the numN values only.
#We will also covert sampleID as a factor stingr and ensure numN,mean_depth and num_mapped_reads are numeric

data1<-data %>% mutate(pass=data$numN<=200)%>% filter(pass)
data2<-data1 %>% mutate(sampleID=as.factor(sampleID), numN=as.numeric(numN), mean_depth=as.numeric(mean_depth),
                        num_mapped_reads= as.numeric(num_mapped_reads),
                        percent_ref_genome_cov=as.numeric(percent_ref_genome_cov))

data2<-data2%>%arrange(numN)
```



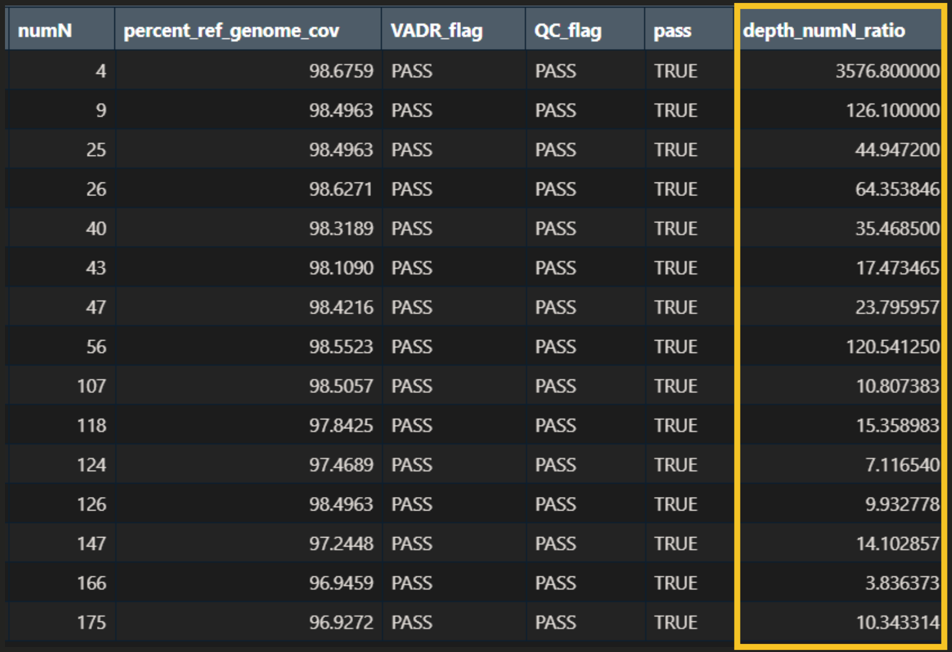
Seotype	Kraken2_Viral_Broad_Percentage	reference	start	end	num_raw_reads	num_clean_reads	num_mapped_reads	percent_mapped_clean_reads	cov_bases_mapped	percent_genome_cov_map	mean_depth	mean_base_qual	mean_map_qual	assembly_length	numN	percent_ref_genome_cov	VADR_flag	QC_flag	pass
4	97.24	NC_002640.1	1	10649	3698680	3131044	942714	30.1086	10510	98.6947	14307.200	37.5	60.0	10512	4	98.6759	PASS	PASS	TRUE
3	99.18	NC_001475.2	1	10707	120412	94536	79817	84.4303	10555	98.5804	1134.900	37.4	60.0	10555	9	98.4963	PASS	PASS	TRUE
3	99.60	NC_001475.2	1	10707	114006	87730	76858	87.6074	10571	98.7298	1123.680	37.4	60.0	10571	25	98.4963	PASS	PASS	TRUE
3	99.87	NC_001475.2	1	10707	179998	135548	111465	82.2329	10586	98.8699	1673.200	37.4	60.0	10586	26	98.6271	PASS	PASS	TRUE
3	99.77	NC_001475.2	1	10707	222878	133620	107503	80.4543	10567	98.6924	1418.740	37.3	60.0	10567	40	98.3189	PASS	PASS	TRUE
1	99.73	NC_001477.1	1	10735	91656	64174	53936	84.0465	10575	98.5095	751.359	37.4	59.8	10575	43	98.1090	PASS	PASS	TRUE
3	99.67	NC_001475.2	1	10707	174796	111808	82800	74.0555	10581	98.8232	1118.410	37.4	60.0	10585	47	98.4216	PASS	PASS	TRUE
3	99.77	NC_001475.2	1	10707	1914544	1453602	474162	32.6198	10608	99.0754	6750.310	37.4	60.0	10608	56	98.5523	PASS	PASS	TRUE
3	99.66	NC_001475.2	1	10707	132492	93254	79987	85.7733	10654	99.5050	1156.390	37.4	60.0	10654	107	98.5057	PASS	PASS	TRUE
3	74.52	NC_001475.2	1	10707	276218	229628	121451	52.8903	10594	98.9446	1812.360	37.5	60.0	10594	118	97.8425	PASS	PASS	TRUE
3	99.74	NC_001475.2	1	10707	157644	81856	67410	82.3519	10560	98.6271	882.451	37.4	60.0	10560	124	97.4689	PASS	PASS	TRUE
3	99.83	NC_001475.2	1	10707	125722	98502	84705	85.9932	10649	99.4583	1251.530	37.4	60.0	10672	126	98.4963	PASS	PASS	TRUE
3	95.55	NC_001475.2	1	10707	412068	330900	153041	46.2499	10559	98.6177	2073.120	37.4	60.0	10559	147	97.2448	PASS	PASS	TRUE
3	99.33	NC_001475.2	1	10707	125442	60842	50781	83.4637	10546	98.4963	636.838	37.3	60.0	10546	166	96.9459	PASS	PASS	TRUE
3	99.26	NC_001475.2	1	10707	291252	250404	127973	51.1066	10553	98.5617	1810.080	37.5	60.0	10553	175	96.9272	PASS	PASS	TRUE

Advanced Molecular Detection
Southeast Region Bioinformatics

❸ Understand



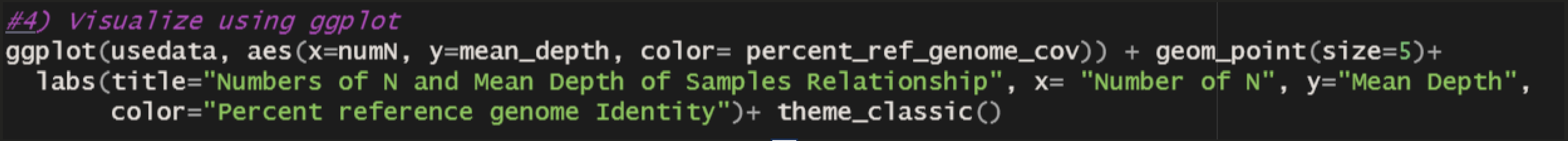
```
#3) Transform
#Use dplyr package to create numN and mean_depth ratio
usedata <- data2 %>%mutate(depth_numN_ratio=mean_depth/numN)
```



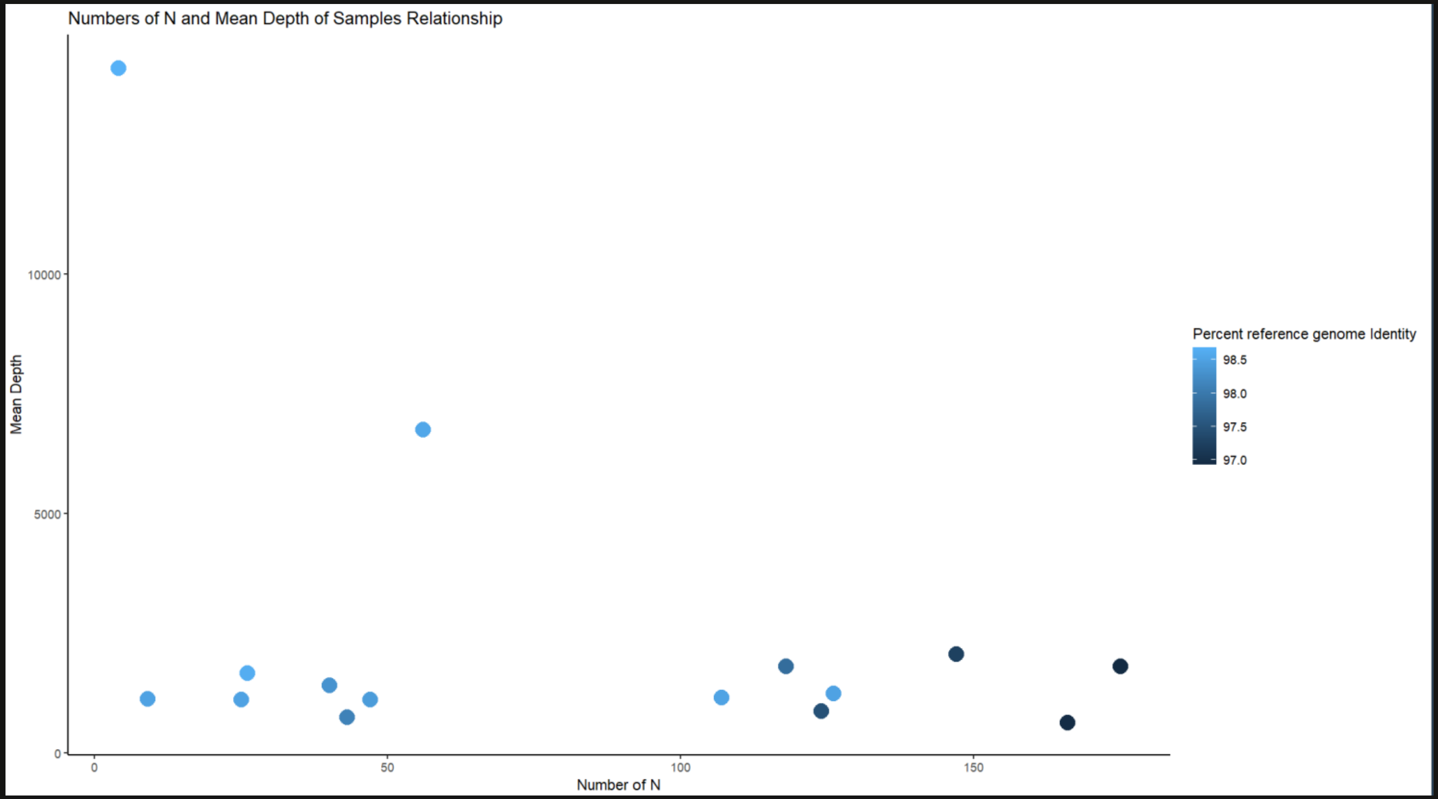
numN	percent_ref_genome_cov	VADR_flag	QC_flag	pass	depth_numN_ratio
4	98.6759	PASS	PASS	TRUE	3576.800000
9	98.4963	PASS	PASS	TRUE	126.100000
25	98.4963	PASS	PASS	TRUE	44.947200
26	98.6271	PASS	PASS	TRUE	64.353846
40	98.3189	PASS	PASS	TRUE	35.468500
43	98.1090	PASS	PASS	TRUE	17.473465
47	98.4216	PASS	PASS	TRUE	23.795957
56	98.5523	PASS	PASS	TRUE	120.541250
107	98.5057	PASS	PASS	TRUE	10.807383
118	97.8425	PASS	PASS	TRUE	15.358983
124	97.4689	PASS	PASS	TRUE	7.116540
126	98.4963	PASS	PASS	TRUE	9.932778
147	97.2448	PASS	PASS	TRUE	14.102857
166	96.9459	PASS	PASS	TRUE	3.836373
175	96.9272	PASS	PASS	TRUE	10.343314

Advanced Molecular Detection
Southeast Region Bioinformatics

❸ Understand Cont.

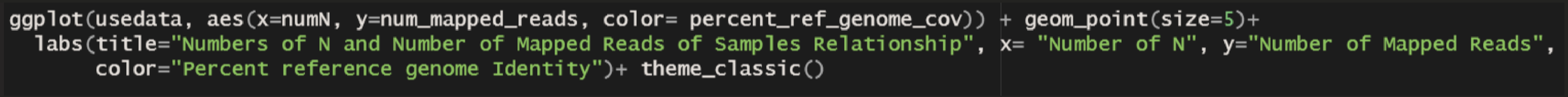


```
#4) Visualize using ggplot
ggplot(usedata, aes(x=numN, y=mean_depth, color= percent_ref_genome_cov)) + geom_point(size=5)+
  labs(title="Numbers of N and Mean Depth of Samples Relationship", x= "Number of N", y="Mean Depth",
       color="Percent reference genome Identity")+ theme_classic()
```

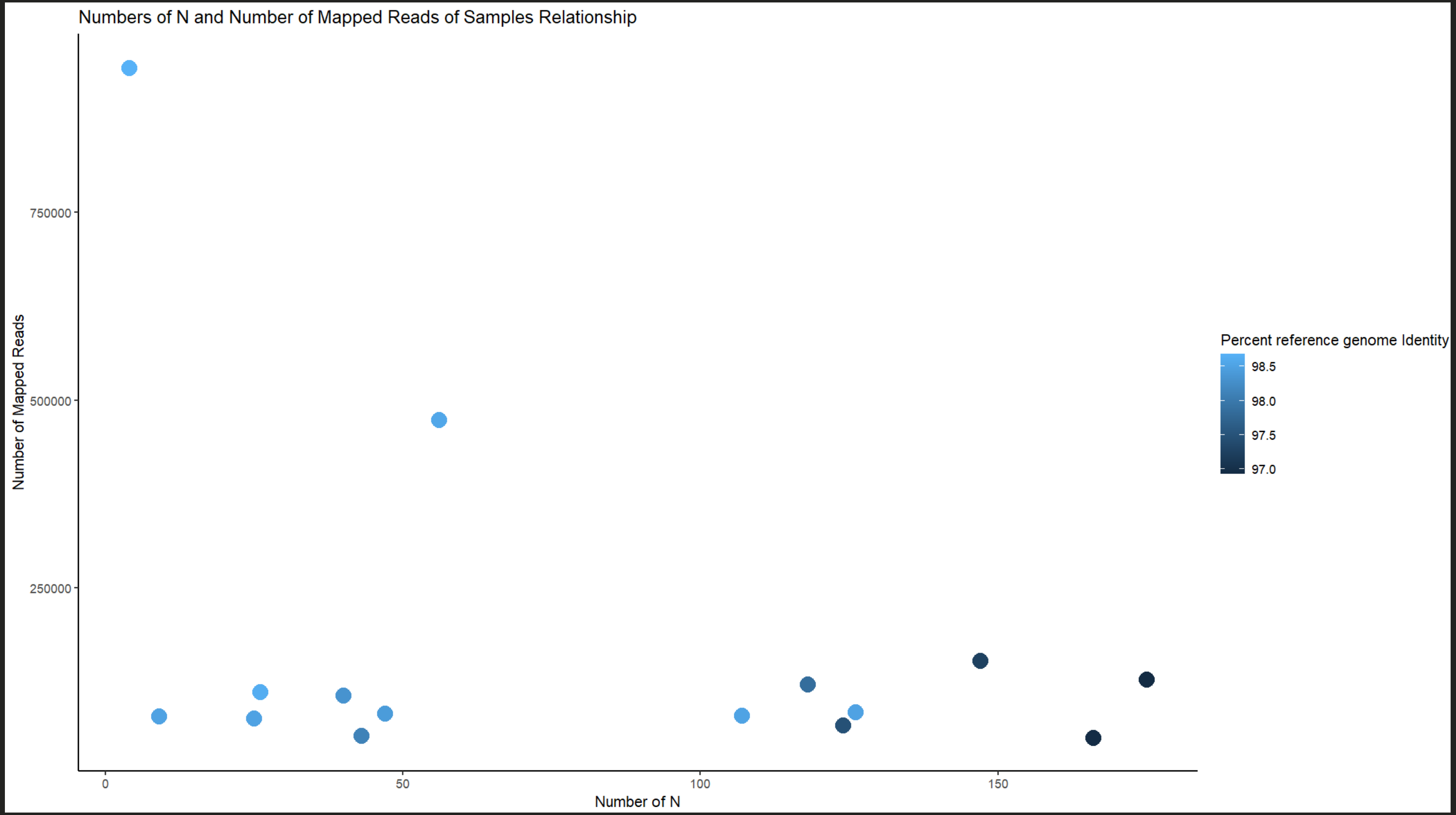


Advanced Molecular Detection
Southeast Region Bioinformatics

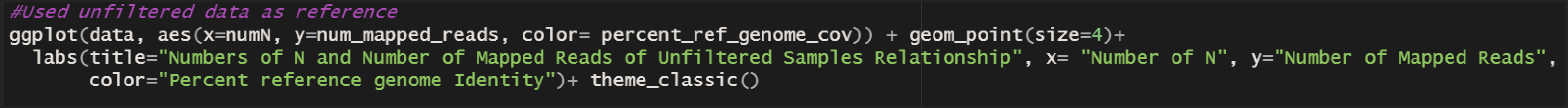
❸ Understand Cont.



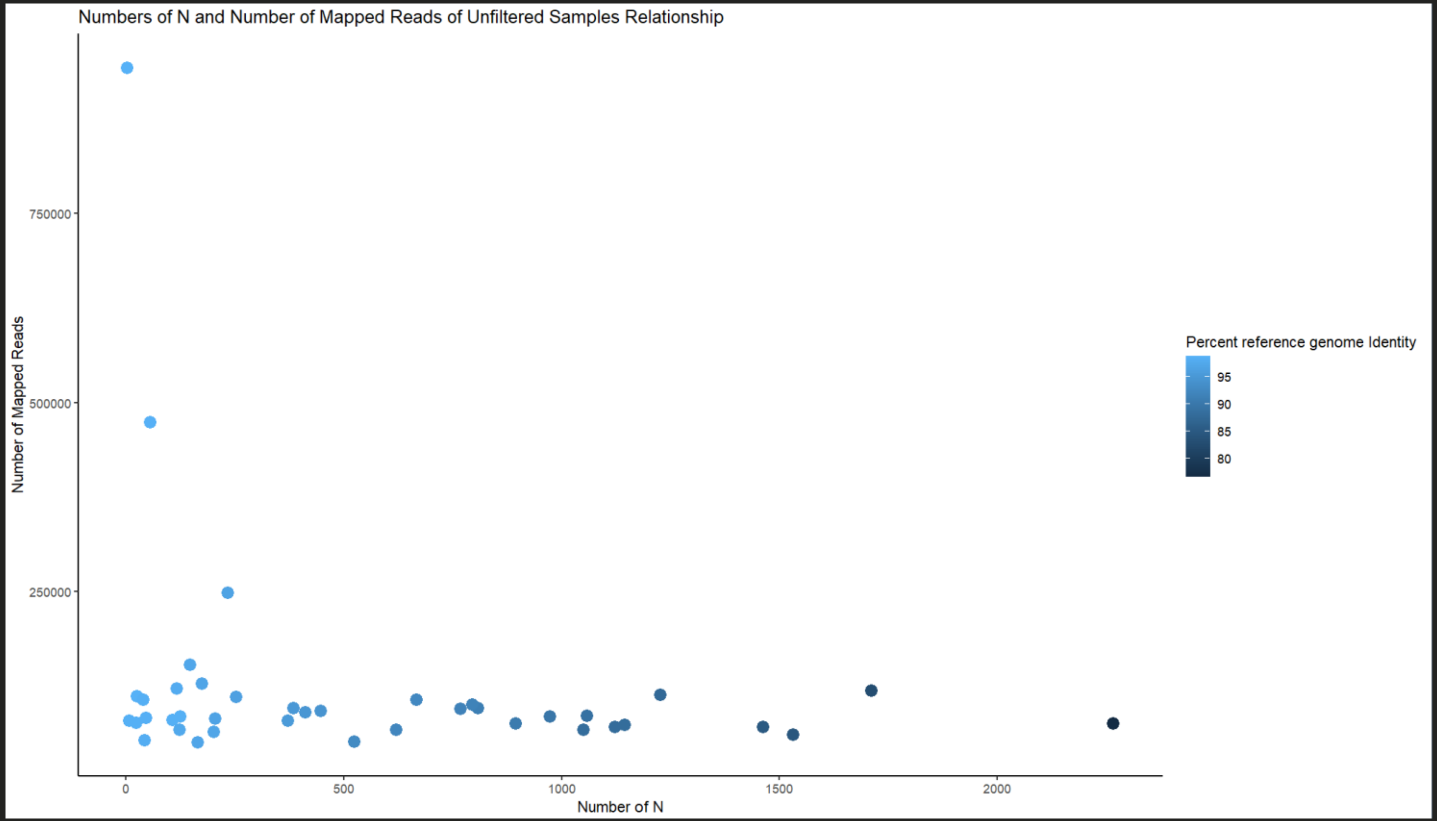
```
ggplot(usedata, aes(x=numN, y=num_mapped_reads, color= percent_ref_genome_cov)) + geom_point(size=5)+
  labs(title="Numbers of N and Number of Mapped Reads of Samples Relationship", x= "Number of N", y="Number of Mapped Reads",
       color="Percent reference genome Identity")+ theme_classic()
```



❸ Understand Cont.



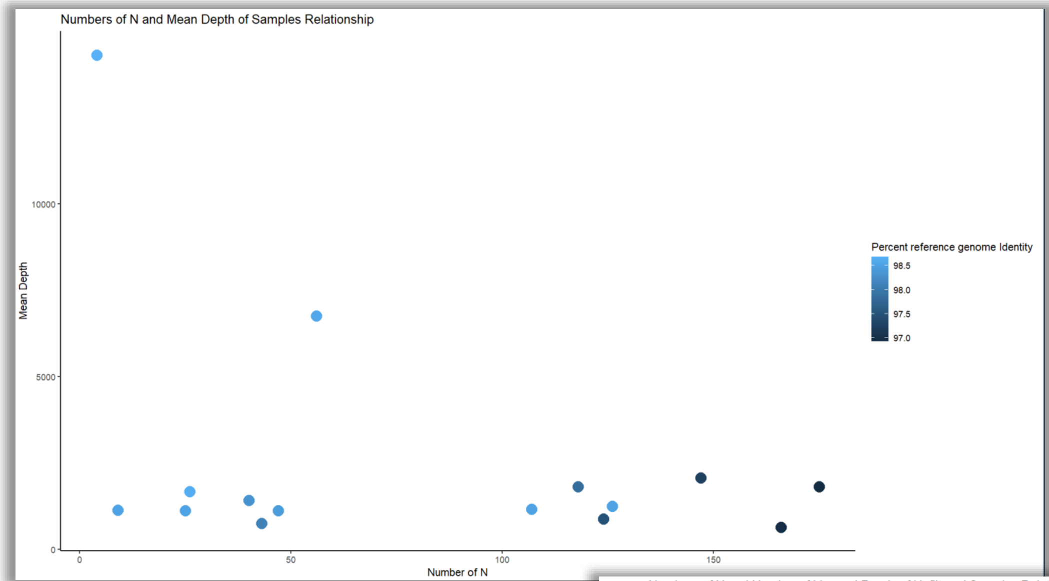
```
#Used unfiltered data as reference
ggplot(data, aes(x=numN, y=num_mapped_reads, color= percent_ref_genome_cov)) + geom_point(size=4)+
  labs(title="Numbers of N and Number of Mapped Reads of Unfiltered Samples Relationship", x= "Number of N", y="Number of Mapped Reads",
       color="Percent reference genome Identity")+ theme_classic()
```

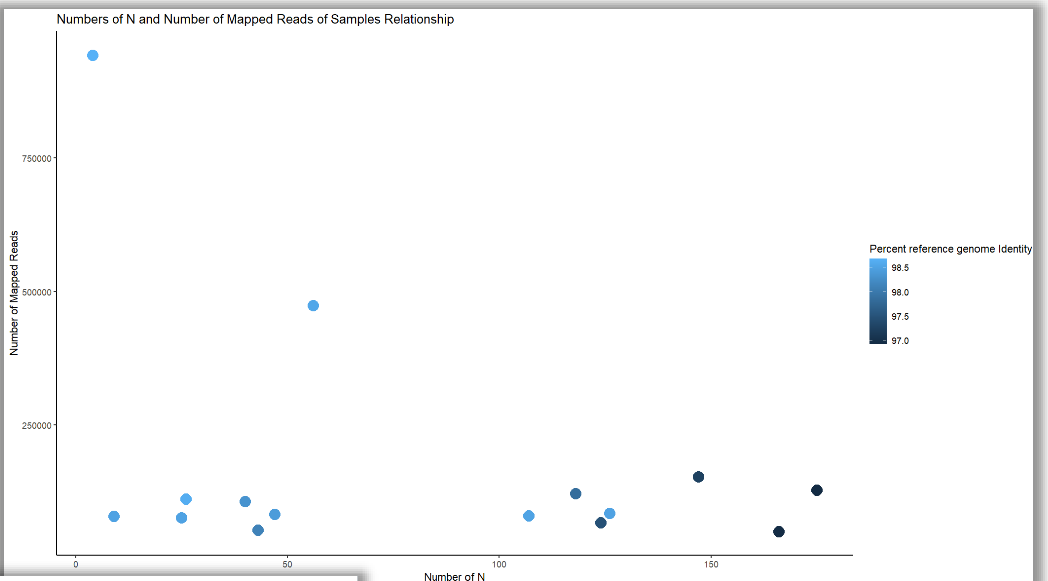


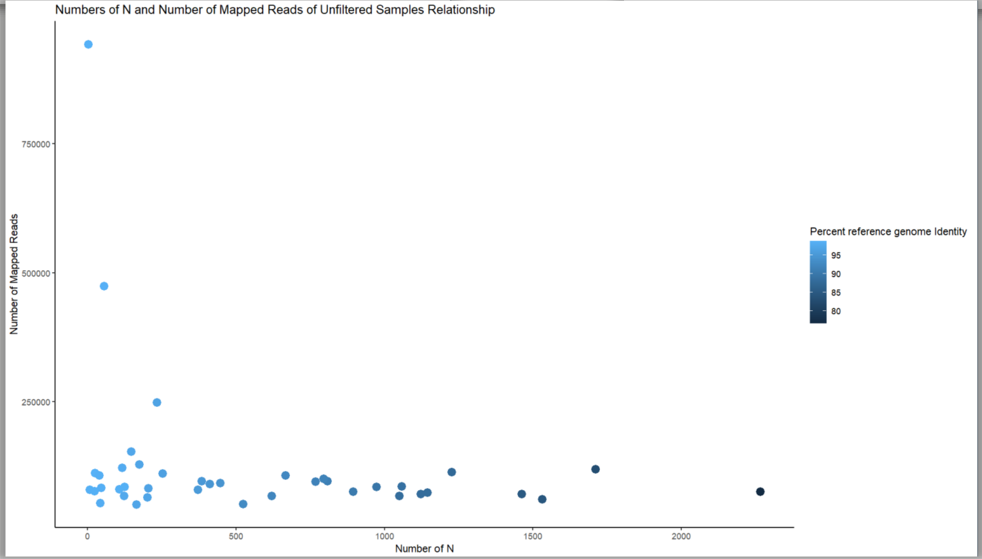
Numbers of N and Number of Mapped Reads of Unfiltered Samples Relationship

Advanced Molecular Detection
Southeast Region Bioinformatics

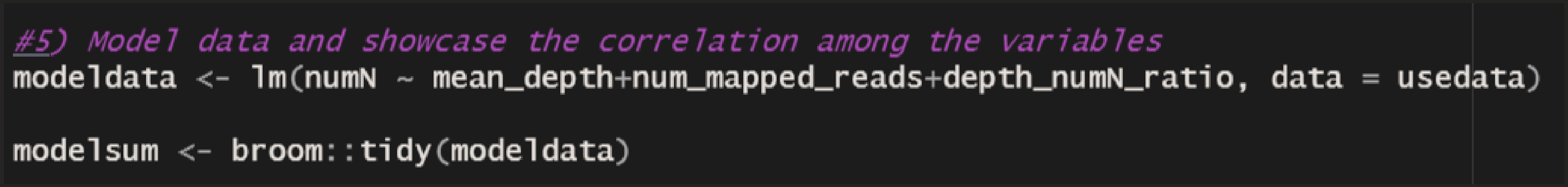
❸ Understand Cont.


Numbers of N and Mean Depth of Samples Relationship

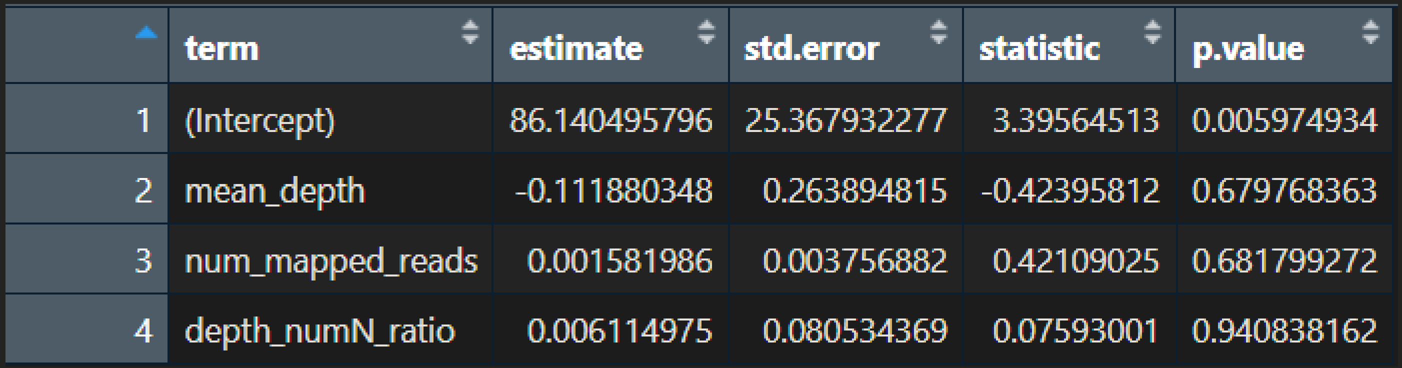

Numbers of N and Number of Mapped Reads of Samples Relationship


Numbers of N and Number of Mapped Reads of Unfiltered Samples Relationship

❸ Understand Cont.



```
#5) Model data and showcase the correlation among the variables
modeldata <- lm(numN ~ mean_depth+num_mapped_reads+depth_numN_ratio, data = usedata)

modelsum <- broom::tidy(modeldata)
```



	term	estimate	std.error	statistic	p.value
1	(Intercept)	86.140495796	25.367932277	3.39564513	0.005974934
2	mean_depth	-0.111880348	0.263894815	-0.42395812	0.679768363
3	num_mapped_reads	0.001581986	0.003756882	0.42109025	0.681799272
4	depth_numN_ratio	0.006114975	0.080534369	0.07593001	0.940838162

Advanced Molecular Detection
Southeast Region Bioinformatics

Conclusion

- ✓ Understood the collection of the packages and how to use the core packages
- ✓ Usage of the data analysis workflow that tidyverse was built upon
- ✓ Experiment that was shown resulted in numN had no significant effect on the mapped reads and mean depth.

Citation

- https://www.rdocumentation.org/packages/tidyverse/versions/2.0.0
- https://ggplot2.tidyverse.org/
- https://r-graph-gallery.com/ggplot2-package.html

Advanced Molecular Detection

Southeast Region Bioinformatics

Questions?

bphl-sebioinformatics@flhealth.gov

Molly Mitchell, PhD
Bioinformatician
Molly.Mitchell@flhealth.gov

Nikhil Reddy, MS
Bioinformatician
Nikhil.Yengala@flhealth.gov

Sam Marcellus, MPH
Bioinformatician
Samantha.marcellus@flhealth.gov