

```
In [2]: import hail as hl

# hl.init(backend="spark")

# Say, "the default sans-serif font is COMIC SANS"
# matplotlib.rcParams['font.sans-serif'] = "Arial"
# Then, "ALWAYS use sans-serif fonts"
# matplotlib.rcParams['font.family'] = "sans-serif"
```

```
In [3]: # import holoviews as hv
import numpy as np
import pandas as pd

from bokeh.layouts import column, gridplot, Plot, row
from bokeh.models import *
from bokeh.plotting import *
from bokeh.palettes import Category10, Spectral6
from gnomad import *
from gnomad.sample_qc.ancestry import POP_COLORS
from gnomad.utils.filtering import add_filters_expr
from gnomad.utils.plotting import *
# from gnomad_qc.v3.resources import *
from IPython.core.display import display, HTML
from statsmodels.robust.scale import mad

# hv.extension("bokeh")

TOOLS = "hover,save,pan,box_zoom,reset,wheel_zoom"
output_notebook()
display(HTML("<style>.container { width:100% !important; }</style>"))

if 'old_show' not in dir():
    old_show = hl.Table.show
    def new_show(t, n=10, width=90, truncate=None, types=True):
        old_show(t, 10, 170, 40)
    hl.Table.show = new_show
```

```
/tmp/ipykernel_52558/3819452698.py:14: DeprecationWarning: Importing display
from IPython.core.display is deprecated since IPython 7.14, please import fro
m IPython display
from IPython.core.display import display, HTML
```



BokehJS 3.1.1 successfully loaded.

```
In [4]: rgp_path = 'gs://marten-seqr-sandbox-storage/rgp_entire_mt_from_vcf_20240404
rgp_mt = hl.read_matrix_table(rgp_path)
```

```

Initializing Hail with default parameters...
/opt/conda/miniconda3/lib/python3.10/site-packages/hailtop/aiocloud/aiogoogle
e/user_config.py:43: UserWarning: Reading spark-defaults.conf to determine GC
S requester pays configuration. This is deprecated. Please use `hailctl confi
g set gcs_requester_pays/project` and `hailctl config set gcs_requester_pays/
buckets`.
  warnings.warn(
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogL
evel(newLevel).
SPARKMONITOR_LISTENER: Started SparkListener for Jupyter Notebook
SPARKMONITOR_LISTENER: Port obtained from environment: 37269
SPARKMONITOR_LISTENER: Application Started: application_1712844629510_0004
...Start Time: 1712864749432

Running on Apache Spark version 3.3.0
SparkUI available at http://dmdragen-m.us-central1-b.c.marten-seqr-sandbox-91
23.internal:41309
Welcome to

      <>
    / / / / _ _ / /
   / _ _ / _ _ / / /
  / / / / \ _ _ / / /   version 0.2.120-f00f916faf78
LOGGING: writing to /home/hail/hail-20240411-1945-0.2.120-f00f916faf78.log

```

```
In [5]: dragen_path = 'gs://marten-seqr-sandbox-storage/dragen_entire_mt_from_vcf_20
dragen_mt = hl.read_matrix_table(dragen_path)
```

```
In [6]: # Hana queries:
# 1: Remove the X chromosome - Stephanie mentioned that in GATK males
# -> on the X chromosome have a ton of GQs between 30 and 40m so that might
# 2: Look at the GQ distribution for non-ref calls only, to see if tehres ar
```

```
In [7]: sex_set = hl.set({"chrX", "chrY"})
```

```
In [8]: dragen_mt_autosomes = dragen_mt.filter_rows(~sex_set.contains(dragen_mt.locu
```

```
In [9]: dragen_mt_nonref = dragen_mt.filter_entries(dragen_mt.GT.is_non_ref())
```

```
In [10]: # GQ distribution: overall vs autosomes vs nonref
```

```
In [11]: hail_expr_set = {'All':dragen_mt.GQ , 'Autosomes':dragen_mt_autosomes.GQ , '
```

```
In [12]: # for hail_title,hail_expr in hail_expr_set.items():
#   plotting_obj = hl.plot.histogram(hail_expr,range=(0,100))
#   plotting_obj.title = f"GQ of DRAGEN {hail_title}"
#   plotting_obj.title.text_font_size = '18pt'
#   show(plotting_obj)
```

```
In [13]: # does this behavior replicate in the prior GATK/WARP callset ?
```

```
In [14]: # rgp_mt
rgp_mt_autosomes = rgp_mt.filter_rows(~sex_set.contains(rgp_mt.locus.contig)
```

```
rgp_mt_nonref = rgp_mt.filter_entries(rgp_mt.GT.is_non_ref())
```

```
In [15]: hail_rgp_expr_set = {'All':rgp_mt.GQ , 'Autosomes':rgp_mt_autosomes.GQ , 'No
```

```
In [16]: # for hail_title,hail_expr in hail_rgp_expr_set.items():
#         plotting_obj = hl.plot.histogram(hail_expr,range=(0,100))
#         plotting_obj.title = f"GQ of GATK/WARP {hail_title}"
#         plotting_obj.title.text_font_size = '18pt'
#         show(plotting_obj)
```

```
In [17]: # for hail_title,hail_expr in hail_rgp_expr_set.items():
#         plotting_obj = hl.plot.histogram(hail_expr,range=(0,100),log=True)
#         plotting_obj.title = f"LOG: GQ of GATK/WARP {hail_title}"
#         plotting_obj.title.text_font_size = '18pt'
#         show(plotting_obj)
```

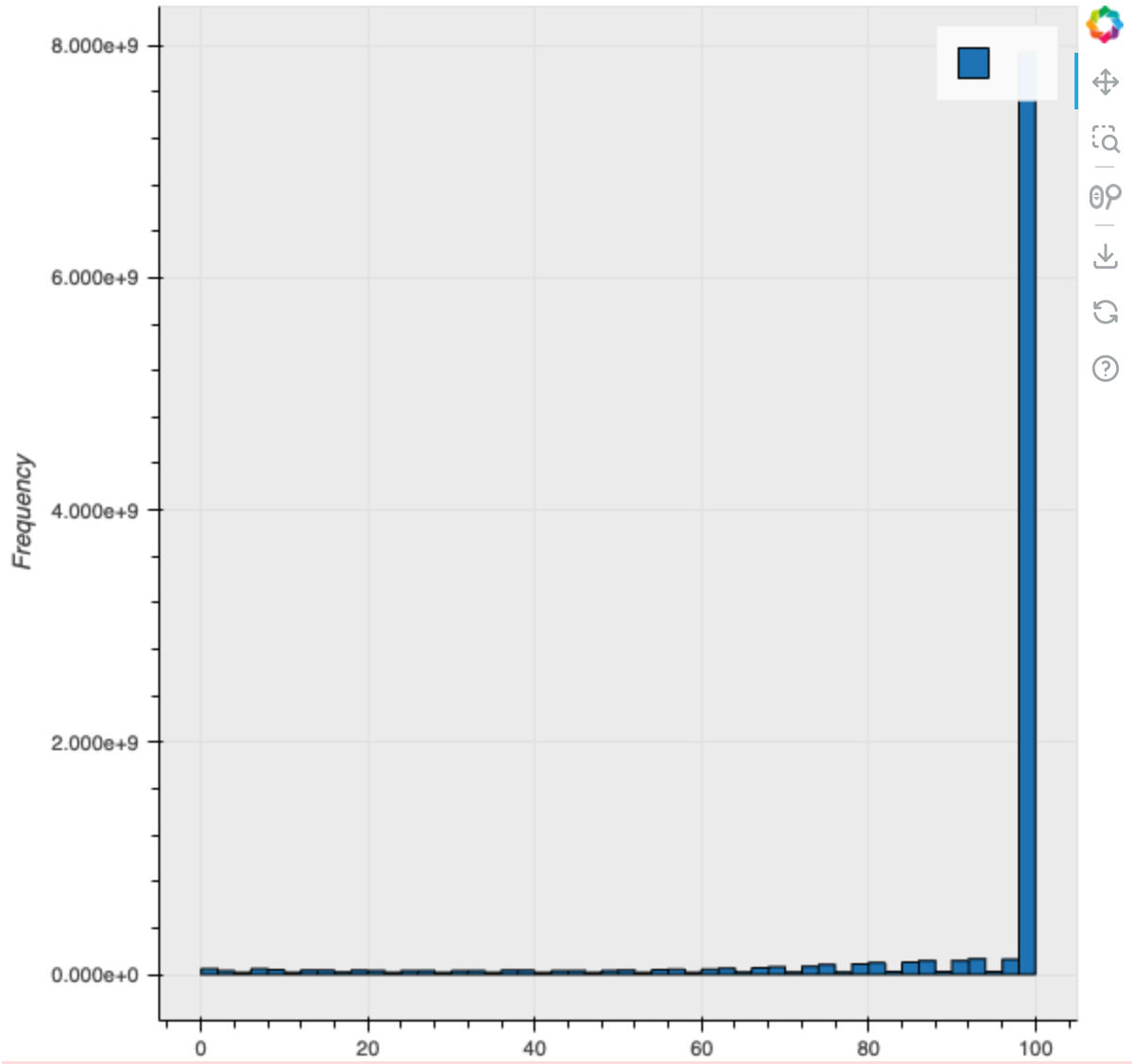
```
In [18]: # Autosomal nonref
dragen_autosomal_nonref = dragen_mt_autosomes.filter_entries(dragen_mt_autos
rgp_autosomal_nonref = rgp_mt_autosomes.filter_entries(rgp_mt_autosomes.GT.i
```

```
In [19]: autosomal_nonref = {"GATK/WARP RGP":rgp_autosomal_nonref.GQ , "DRAGEN/GVS":c
```

```
In [20]: for hail_title,hail_expr in autosomal_nonref.items():
plotting_obj = hl.plot.histogram(hail_expr,range=(0,100))
plotting_obj.title = f"GQ of Autosomal Non-Ref {hail_title}"
plotting_obj.title.text_font_size = '18pt'
show(plotting_obj)
```

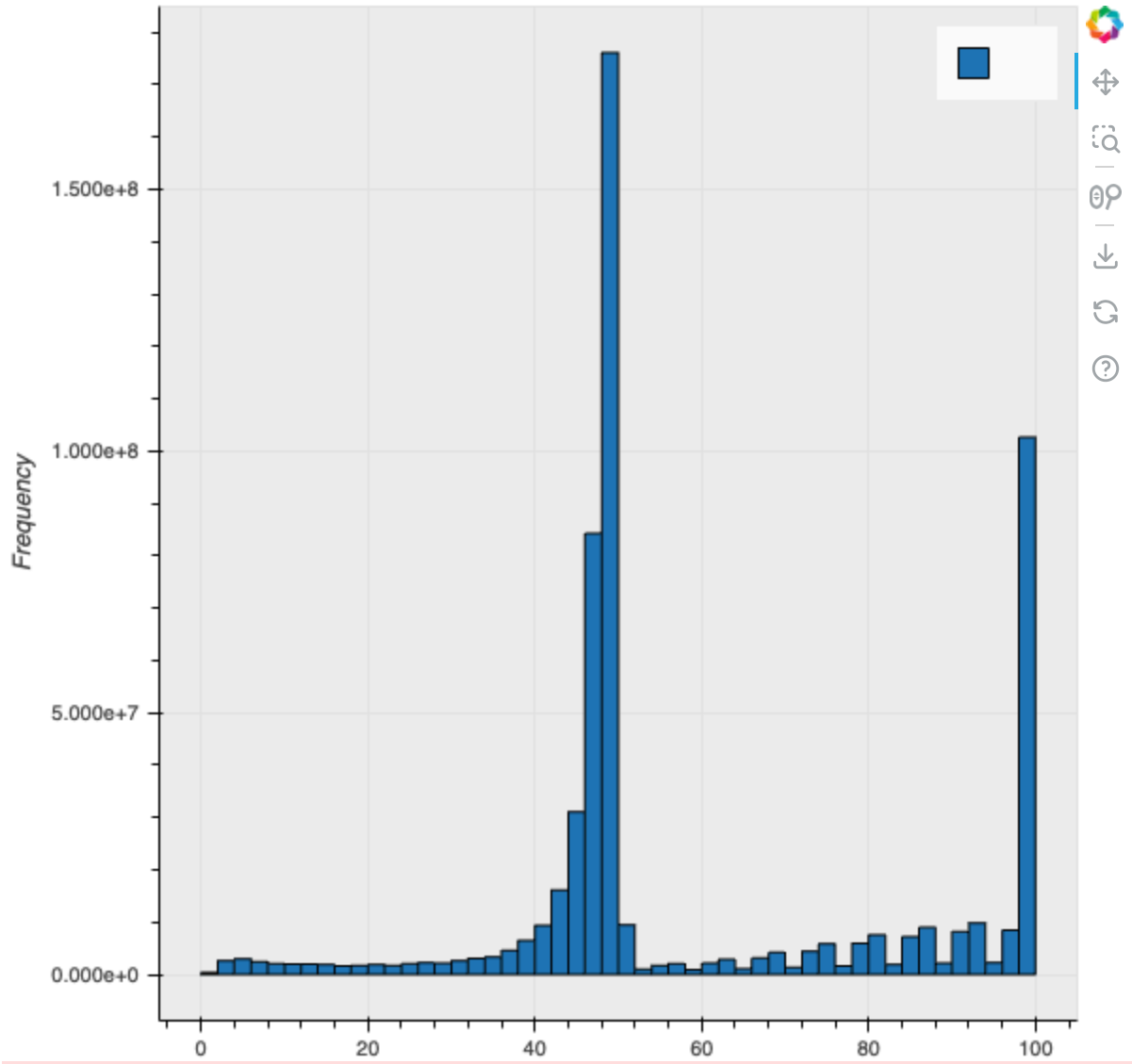
```
[Stage 1:=====> (98 + 5) / 100]
```

GQ of Autosomal Non-Ref GATK/WARP RGP



[Stage 3: ===== (510 + 2) / 5
10]

GQ of Autosomal Non-Ref DRAGEN/GVS



[Stage 3: =====(510 + 1) / 5
10]

In []: