In [1]:
```
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 [2]:
```
# 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_9044/3819452698.py:14: DeprecationWarning: Importing display f
rom IPython.core.display is deprecated since IPython 7.14, please import from
IPython display
  from IPython.core.display import display, HTML
```

BokehJS 3.1.1 successfully loaded.

In [3]:
```
dragen_called_rgp = hl.read_matrix_table('gs://marten-seqr-sandbox-storage/c
```

```
Initializing Hail with default parameters...
/opt/conda/miniconda3/lib/python3.10/site-packages/hailtop/aiocloud/aiogoogl
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: 43987
SPARKMONITOR_LISTENER: Application Started: application_1712844629510_0001
...Start Time: 1712845513125
Running on Apache Spark version 3.3.0
SparkUI available at http://dmdragen-m.us-central1-b.c.marten-seqr-sandbox-91
23.internal:41293
Welcome to
     __  __     <>__
    / /_/ /__  __/ /
   / __  / _ `/ / /
  /_/ /_/\_,_/_/_/   version 0.2.120-f00f916faf78
LOGGING: writing to /home/hail/hail-20240411-1425-0.2.120-f00f916faf78.log
```

In [4]:
```
gatk_called_rgp = hl.read_matrix_table('gs://marten-seqr-sandbox-storage/gat
```

In [5]:
```
# dragen_massive_mt = hl.import_vcf('gs://seqr-scratch-temp/gregor_consortiu
#                                   reference_genome="GRCh38",
#                                   force_bgz=True,
#                                   array_elements_required=False)
```

In [6]:
```
# dragen_massive_mt = dragen_massive_mt.repartition(1500)
```

In [7]:
```
dragen_massive_mt = hl.read_matrix_table('gs://seqr-scratch-temp/gregor_cons
```

In [8]:
```
dragen_massive_mt.describe()
```

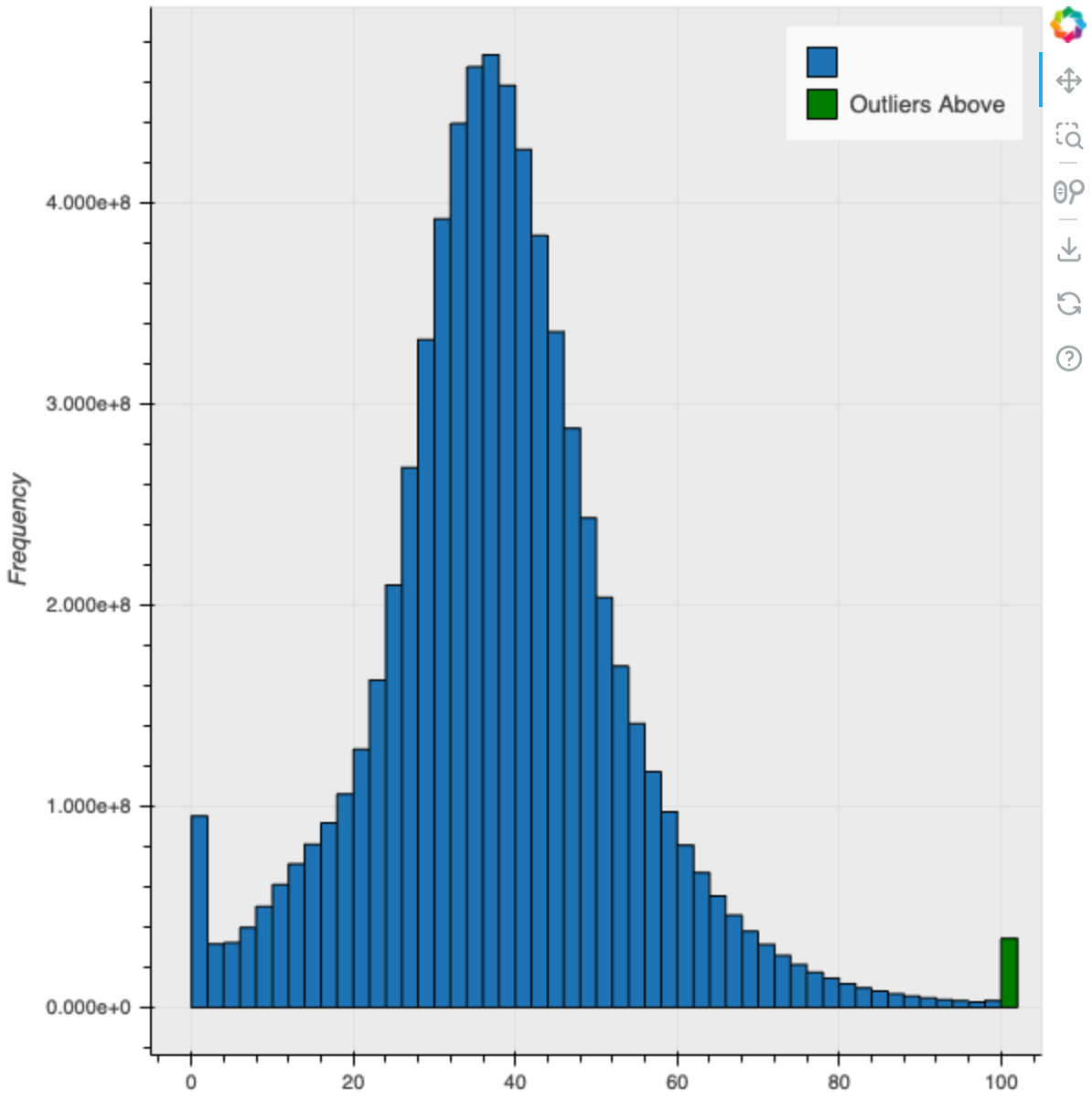
```
----------------------------------------
Global fields:
    None
----------------------------------------
Column fields:
    's': str
----------------------------------------
Row fields:
    'locus': locus<GRCh38>
    'alleles': array<str>
    'rsid': str
    'qual': float64
    'filters': set<str>
    'info': struct {
        AC: array<int32>,
        AF: array<float64>,
        AN: int32,
        AS_QUALapprox: str,
        AS_YNG: array<str>,
        CALIBRATION_SENSITIVITY: array<str>,
        QUALapprox: int32,
        SCORE: array<str>
    }
----------------------------------------
Entry fields:
    'AD': array<int32>
    'FT': str
    'GQ': int32
    'GT': call
    'RGQ': int32
----------------------------------------
Column key: ['s']
Row key: ['locus', 'alleles']
----------------------------------------
```

In [9]:
```
# Check Coverage
dragen_massive_mt = dragen_massive_mt.annotate_entries(
    total_cov = hl.if_else(~hl.is_missing(dragen_massive_mt.AD),dragen_massi
)
```

In [13]:
```
total_cov_dragen_obj = hl.plot.histogram(dragen_massive_mt.total_cov,range=(
```

```
[Stage 5:=====================================================>   (28 + 5) /
30]
```

In [14]:
```
show(total_cov_dragen_obj)
```



In [11]:
```
dragen_massive_mt.total_cov.summarize()
```

```
[Stage 3:=================================================>(1499 + 1) / 1500]
```

95411989640 records.

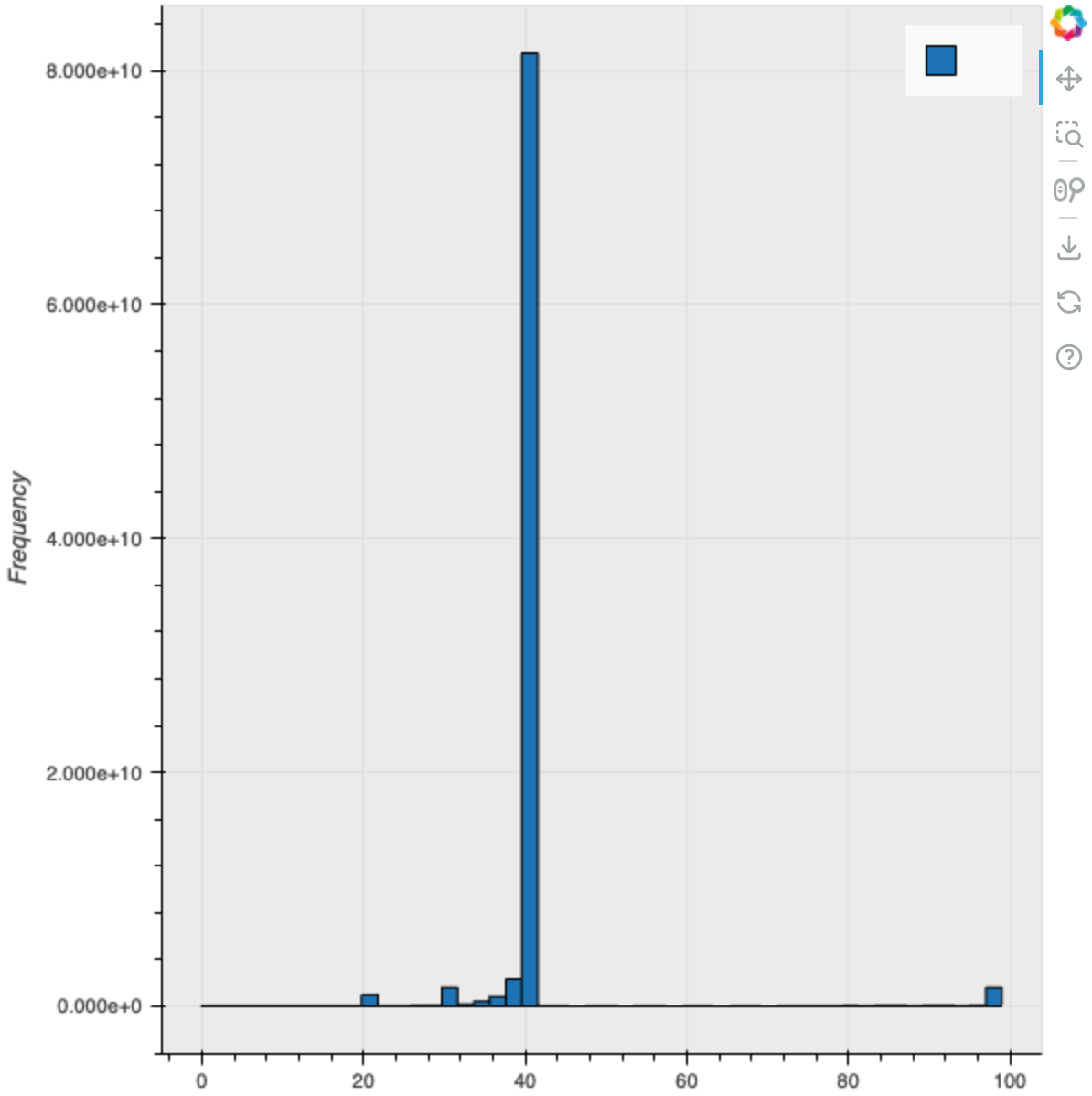
total_cov (*float64*):

Non-missing	6886429869 (7.22%)
Missing	88525559771 (92.78%)
Minimum	0.00
Maximum	13190.00
Mean	38.39
Std Dev	38.10

In [12]:
```
# Check GQ Distribution
```

In [15]:
```
show(hl.plot.histogram(dragen_massive_mt.GQ))
```

```
[Stage 8:================================================>       (26 + 8) / 30]
```



In [16]:
```
dragen_massive_mt.GQ.summarize()
```

```
[Stage 9:=========================================================>(1499 + 1) / 1500]
```

95411989640 records.

GQ (*int32*):

Non-missing	90705470249 (95.07%)
Missing	4706519391 (4.93%)
Minimum	0
Maximum	99
Mean	40.79
Std Dev	9.24

In [17]:
```
# Check ploidy in sex chromosomes
```

In [18]:
```
dragen_massive_mt_autosomes = dragen_massive_mt.filter_rows((dragen_massive_
dragen_massive_mt_xy = dragen_massive_mt.filter_rows((dragen_massive_mt.locu

# mt = mt.filter_rows(mt.locus.contig != 'Y')
```

In [21]:
```
dragen_massive_mt_xy = dragen_massive_mt_xy.annotate_entries(
    diploidy = dragen_massive_mt_xy.GT.is_diploid()
)
```

In [23]:
```
dragen_massive_mt_xy.aggregate_entries(hl.agg.counter(dragen_massive_mt_xy.c
```

```
[Stage 11:=====================================================(1500 + 1) / 15
00]
```

Out[23]:
```
{True: 4062209898, None: 522987142}
```

In [26]:
```
dragen_massive_mt_xy.entries().head(100).show(100)
```

```
[Stage 25:=====================================================>(1428 + 2) / 14
29]
```

					info	
locus	alleles	rsid	qual	filters	AC	AF
locus<GRCh38>	array<str>	str	float64	set<str>	array<int32>	array<
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e

					info	
locus	alleles	rsid	qual	filters	AC	AF
locus<GRCh38>	array<str>	str	float64	set<str>	array<int32>	array<
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e

					info	
locus	alleles	rsid	qual	filters	AC	AF
locus<GRCh38>	array<str>	str	float64	set<str>	array<int32>	array<
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e·

					info	
locus	alleles	rsid	qual	filters	AC	AF
locus<GRCh38>	array<str>	str	float64	set<str>	array<int32>	array<
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e
chrX:10009	["A","G","C"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}	[6,4]	[7.70e

In [28]:
```
dragen_massive_mt_xy.entries().tail(100).show(100)
```

```
[Stage 30:>                                                         (0 + 1) /
1]
```

locus	alleles	rsid	qual	filters
locus<GRCh38>	array<str>	str	float64	set<str>
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}

locus	alleles	rsid	qual	filters
locus<GRCh38>	array<str>	str	float64	set<str>
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}

locus	alleles	rsid	qual	filters
locus<GRCh38>	array<str>	str	float64	set<str>
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}

locus	alleles	rsid	qual	filters
locus<GRCh38>	array<str>	str	float64	set<str>
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}
chrY:56887898	["T","TCTCATGTGTG","TCTCATG"]	NA	-1.00e+01	{"NO_HQ_GENOTYPES"}

In [27]:
```
dragen_massive_mt_xy.aggregate_rows(hl.agg.counter(dragen_massive_mt_xy.locu
```

```
[Stage 27:=====================================================>   (28 + 2) / 
30]
```

Out[27]:
```
{'chrX': 2912448, 'chrY': 316564}
```

In [29]:
```
# Check Filters and FT fields
```

In [30]:
```
dragen_massive_mt.aggregate_rows(hl.agg.counter(dragen_massive_mt.filters))
```

```
[Stage 31:=====================================================(1500 + 7) / 15
00]
```

Out[30]:
```
{frozenset({'ExcessHet'}): 133601,
 frozenset({'ExcessHet', 'NO_HQ_GENOTYPES'}): 2083,
 frozenset({'LowQual'}): 1027047,
 frozenset({'LowQual', 'NO_HQ_GENOTYPES'}): 3503236,
 frozenset({'NO_HQ_GENOTYPES'}): 1628185,
 None: 60897390}
```

In [31]:
```
dragen_massive_mt.aggregate_entries(hl.agg.counter(dragen_massive_mt.FT))
```

```
[Stage 34:=====================================================>   (28 + 4) / 
30]
```

Out[31]:
```
{'PASS': 3989654553,
 'high_CALIBRATION_SENSITIVITY_INDEL': 30868995,
 'high_CALIBRATION_SENSITIVITY_SNP': 276855112,
 None: 91114610980}
```

In [34]:
```
# Per sample and silliest sample qc (the easy way) work
```

```
[Stage 35:=====================================================(1500 + 1) / 1500]
```

In [33]:
```
dragen_massive_mt.aggregate_rows(hl.agg.counter(dragen_massive_mt.locus.cont
```

```
[Stage 35:=====================================================(1500 + 5) / 1500]
```

Out[33]:
```
{'chr1': 5187855,
 'chr10': 3177805,
 'chr11': 3061825,
 'chr12': 2988988,
 'chr13': 2198114,
 'chr14': 2062402,
 'chr15': 1957285,
 'chr16': 2172205,
 'chr17': 1929749,
 'chr18': 1729495,
 'chr19': 1536421,
 'chr2': 5421256,
 'chr20': 1447284,
 'chr21': 982931,
 'chr22': 1075558,
 'chr3': 4422592,
 'chr4': 4427180,
 'chr5': 4053252,
 'chr6': 3868124,
 'chr7': 3759294,
 'chr8': 3494902,
 'chr9': 3008013,
 'chrX': 2912448,
 'chrY': 316564}
```

In [40]:
```
# dragen_massive_mt_entries = dragen_massive_mt.entries()
```

In [39]:
```
# dragen_massive_mt_entries.aggregate(hl.agg.counter(dragen_massive_mt_entri
```

In [41]:
```
dragen_massive_sqc_cols = hl.sample_qc(dragen_massive_mt).cols()
```

```
2024-04-11 15:29:54.936 Hail: WARN: cols(): Resulting column table is sorted by 'col_key'.
    To preserve matrix table column order, first unkey columns with 'key_cols_by()'
```

In [42]:
```
dragen_massive_sqc_cols.show()
```

```
Exception in thread "Thread-39" java.lang.NullPointerException206 + 128) / 15
00]
        at sparkmonitor.listener.JupyterSparkMonitorListener$TaskUpdaterThrea
d.$anonfun$run$1(CustomListener.scala:116)
        at scala.collection.TraversableLike$grouper$1$.apply(TraversableLike.
scala:465)
        at scala.collection.TraversableLike$grouper$1$.apply(TraversableLike.
scala:455)
        at scala.collection.mutable.ResizableArray.foreach(ResizableArray.sca
la:62)
        at scala.collection.mutable.ResizableArray.foreach$(ResizableArray.sc
ala:55)
        at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:49)
        at scala.collection.TraversableLike.groupBy(TraversableLike.scala:52
4)
        at scala.collection.TraversableLike.groupBy$(TraversableLike.scala:45
4)
        at scala.collection.AbstractTraversable.groupBy(Traversable.scala:10
8)
        at sparkmonitor.listener.JupyterSparkMonitorListener$TaskUpdaterThrea
d.run(CustomListener.scala:116)
        at java.base/java.lang.Thread.run(Thread.java:829)
[Stage 41:=====================================================>   (28 + 4) /
30]
```

sample_qc

gq_stats

s	mean	stdev	min	max	call_rate	n_called	n_not_called	r
str	float64	float64	float64	float64	float64	int64	int64	i
"1069047"	4.10e+01	9.41e+00	0.00e+00	9.90e+01	9.51e-01	63907109	3284433	
"1069048"	4.08e+01	9.41e+00	0.00e+00	9.90e+01	9.52e-01	63932842	3258700	
"1069049"	4.08e+01	8.69e+00	0.00e+00	9.90e+01	9.47e-01	63639301	3552241	
"1069050"	4.06e+01	8.94e+00	0.00e+00	9.90e+01	9.50e-01	63800592	3390950	
"1069051"	4.06e+01	8.93e+00	0.00e+00	9.90e+01	9.49e-01	63793943	3397599	
"1069052"	4.07e+01	9.16e+00	0.00e+00	9.90e+01	9.51e-01	63930782	3260760	
"1069053"	4.07e+01	9.18e+00	0.00e+00	9.90e+01	9.50e-01	63855090	3336452	
"1069054"	4.07e+01	9.01e+00	0.00e+00	9.90e+01	9.51e-01	63880505	3311037	
"1069055"	4.10e+01	9.53e+00	0.00e+00	9.90e+01	9.59e-01	64408528	2783014	
"1069056"	4.10e+01	9.44e+00	0.00e+00	9.90e+01	9.56e-01	64221783	2969759	

showing top 10 rows

In [44]:
```
dragen_massive_sqc_cols.sample_qc.n_called.summarize()
```

```
[Stage 44:=====================================================> (29 + 2) / 30]
```

1420 records.

n_called (*int64*):

Non-missing	1420 (100.00%)
Missing	0
Minimum	52506883
Maximum	64822099
Mean	63877091.72
Std Dev	782755.88

In [45]:
```
dragen_massive_sqc_cols.sample_qc.n_non_ref.summarize()
```

```
[Stage 47:=============================================================(30 + 1) / 30]
```

1420 records.

n_non_ref (*int64*):

Non-missing	1420 (100.00%)
Missing	0
Minimum	4536743
Maximum	5872038
Mean	4849598.50
Std Dev	184463.26

In [47]:
```
dragen_massive_sqc_cols.sample_qc.n_singleton.summarize()
```

```
[Stage 50:=====================================================>   (28 + 4) / 30]
```

1420 records.

n_singleton (*int64*):

Non-missing	1420 (100.00%)
Missing	0
Minimum	1408
Maximum	131434
Mean	17360.43
Std Dev	17187.90

In [48]:
```
# Odd question - so all calls are diploid, but is anything odd on sex chromo
```

In [53]:
```
# dragen_massive_mt_xy.aggregate_rows(hl.agg.counter(dragen_massive_mt_xy.is
```

In [51]:
```
dragen_massive_mt_xy.aggregate_rows(hl.agg.counter(dragen_massive_mt_xy.locu
```

```
[Stage 53:================================================>     (27 + 3) /
30]
```

Out[51]:
```
{False: 3091704, True: 137308}
```

In [52]:
```
dragen_massive_mt_xy.aggregate_rows(hl.agg.counter(dragen_massive_mt_xy.locu
```

```
[Stage 55:=========================================================> (29 + 2) /
30]
```

Out[52]:
```
{False: 3229012}
```

In [54]:
```
dragen_massive_mt_xonly = dragen_massive_mt_xy.filter_rows(dragen_massive_mt

dragen_massive_mt_yonly = dragen_massive_mt_xy.filter_rows(dragen_massive_mt
```

In [55]:
```
xonly_sqc = hl.sample_qc(dragen_massive_mt_xonly).cols()
```

In [56]:
```
yonly_sqc = hl.sample_qc(dragen_massive_mt_yonly).cols()
```

In [57]:
```
xonly_sqc.sample_qc.n_called.show()
```

```
2024-04-11 17:15:19.638 Hail: INFO: Coerced sorted dataset        (13 + 6) /
16]
[Stage 58:==============================================================(16 + 1) /
16]
```

s	<expr>
str	int64
"1069047"	2782242
"1069048"	2685710
"1069049"	2759123
"1069050"	2669154
"1069051"	2665800
"1069052"	2680369
"1069053"	2672281
"1069054"	2675965
"1069055"	2730696
"1069056"	2797563

showing top 10 rows

In [61]:
```
xonly_sqc.sample_qc.n_non_ref.summarize()
```

```
[Stage 69:=============================================================(65 + 1) / 65]
```

1420 records.

n_non_ref (*int64*):

Non-missing	1420 (100.00%)
Missing	0
Minimum	108516
Maximum	226248
Mean	139984.19
Std Dev	23497.89

In [63]:
```
yonly_sqc.sample_qc.n_non_ref.summarize()
```

```
[Stage 72:=================================================>       (7 + 1) / 8]
```

1420 records.

n_non_ref (*int64*):

Non-missing	1420 (100.00%)
Missing	0
Minimum	5725
Maximum	14311
Mean	9727.80
Std Dev	1485.14

In [ ]: