

GHP-2-2 and PAHAT_1 - InDels Analysis

Original Data Extracted from VCF File

```
In [1]: from VCFtoTable import *
from GTtable import *
from GTplots import *
from GTplot import *
from BarPlots import *
from CTbarPlots import *
from variant_hist import*
from stats import *
from FilterVCF import *
from GTfilter import*
from CTguide import *
```

```
In [2]: vcf_sorghum = '/home/anibal/genome_files/freebayes~bwa~GCF_000003195.3_Sorghum_k
```

```
In [3]: samples_all, vcf_df, chrom_len = VCFtoTable(vcf_sorghum)
```

```
In [4]: samples_all
```

```
Out[4]: array(['GHP-2-2', 'PAHAT_1', 'PAHAT_2'], dtype=object)
```

```
In [5]: progenitor = 'PAHAT_1'
mutant = 'GHP-2-2'
samples = [progenitor, mutant]
samples
```

```
Out[5]: ['PAHAT_1', 'GHP-2-2']
```

```
In [6]: vcf_df
```

```
Out[6]:
```

	CHROM	POS	REF	ALT	QUAL	DP	GHP-2-2_DP
0	NC_008360.1	3598	GAAAAAACT	GAAAAAAAAACT	2072.229980	1771	187
1	NC_008360.1	3788		AT	41349.500000	1477	189
2	NC_008360.1	4068		T	38302.101562	1387	177
3	NC_008360.1	7055		G	7958.569824	351	70
4	NC_008360.1	9183		G	246.141006	26	11
...
2276907	NC_012879.2	61191248		GCCA	27.908600	71	16
Processing math: 100%	NC_012879.2	61191286		C	13.749700	70	19

	CHROM	POS	REF	ALT	QUAL	DP	GHP-2-2_DP
2276909	NC_012879.2	61191293	G	C	23.474300	65	17
2276910	NC_012879.2	61233448	GTTCAGGGTTAAGGGT	GTTCAGGGTT	432.449005	139	15
2276911	NC_012879.2	61233618	GTTCAG	GTTCAG	304.154999	117	24

2276912 rows × 17 columns

◀ ▶

In [7]: chrom_len

Out[7]: LEN

CHROM	LEN
NC_012870.2	80884392
NC_012871.2	77742459
NC_012872.2	74386277
NC_012873.2	68658214
NC_012874.2	71854669
NC_012875.2	61277060
NC_012876.2	65505356
NC_012877.2	62686529
NC_012878.2	59416394
NC_012879.2	61233695
NC_008360.1	468628
NC_008602.1	140754

PART 0: Raw

Contingency Table - RAW - All Chromosomes - (No 0/0, 0/1, 1/1 Filtered)

In [8]: contingency_table_0 = contingency_table(samples, vcf_df, 'all')

Contingency Table - Chromosome all

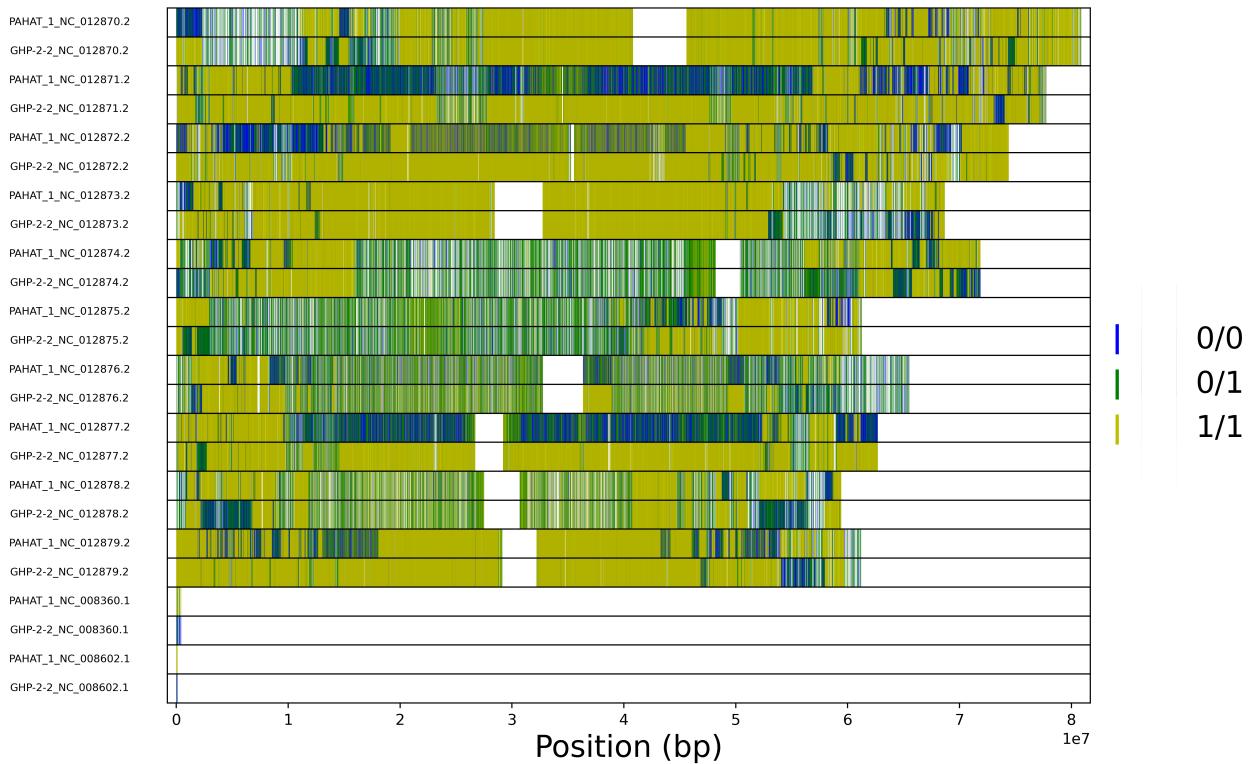
PAHAT_1_GT	GHP-2-2_GT			
	0/0	0/1	1/1	other
0/0	7586	212925	620457	68933
0/1	119419	225957	42168	68933
1/1	307284	54103	618080	68933
other	68933	68933	68933	68933

GT Plot - RAW - All Chromosomes - (No 0/0, 1/1, 'Other' GTs Filtered)

Processing math: 100% close('all')

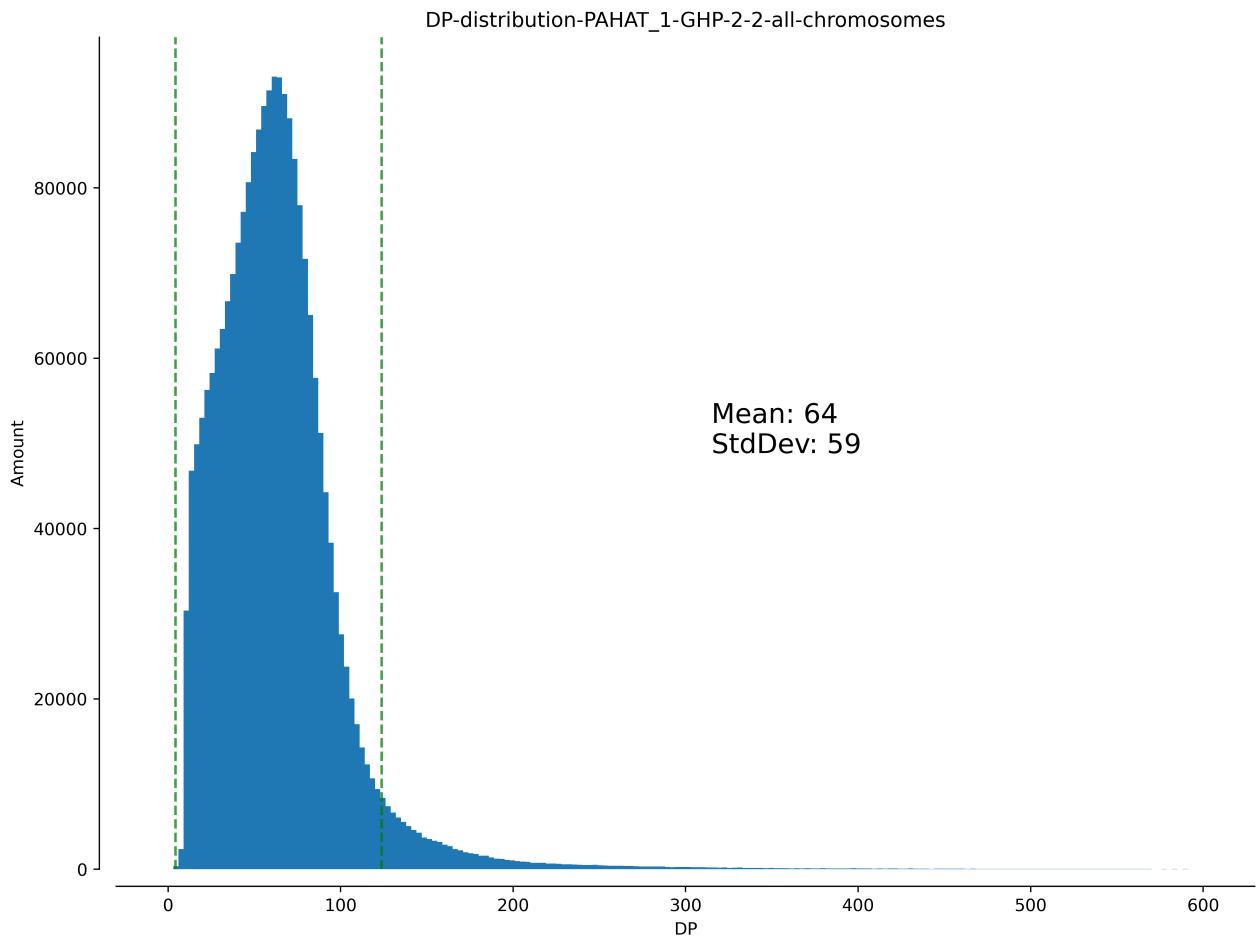
```
GTplot(samples, vcf_df, chrom_len)
```

gt-plot-PAHAT_1-GHP-2-2



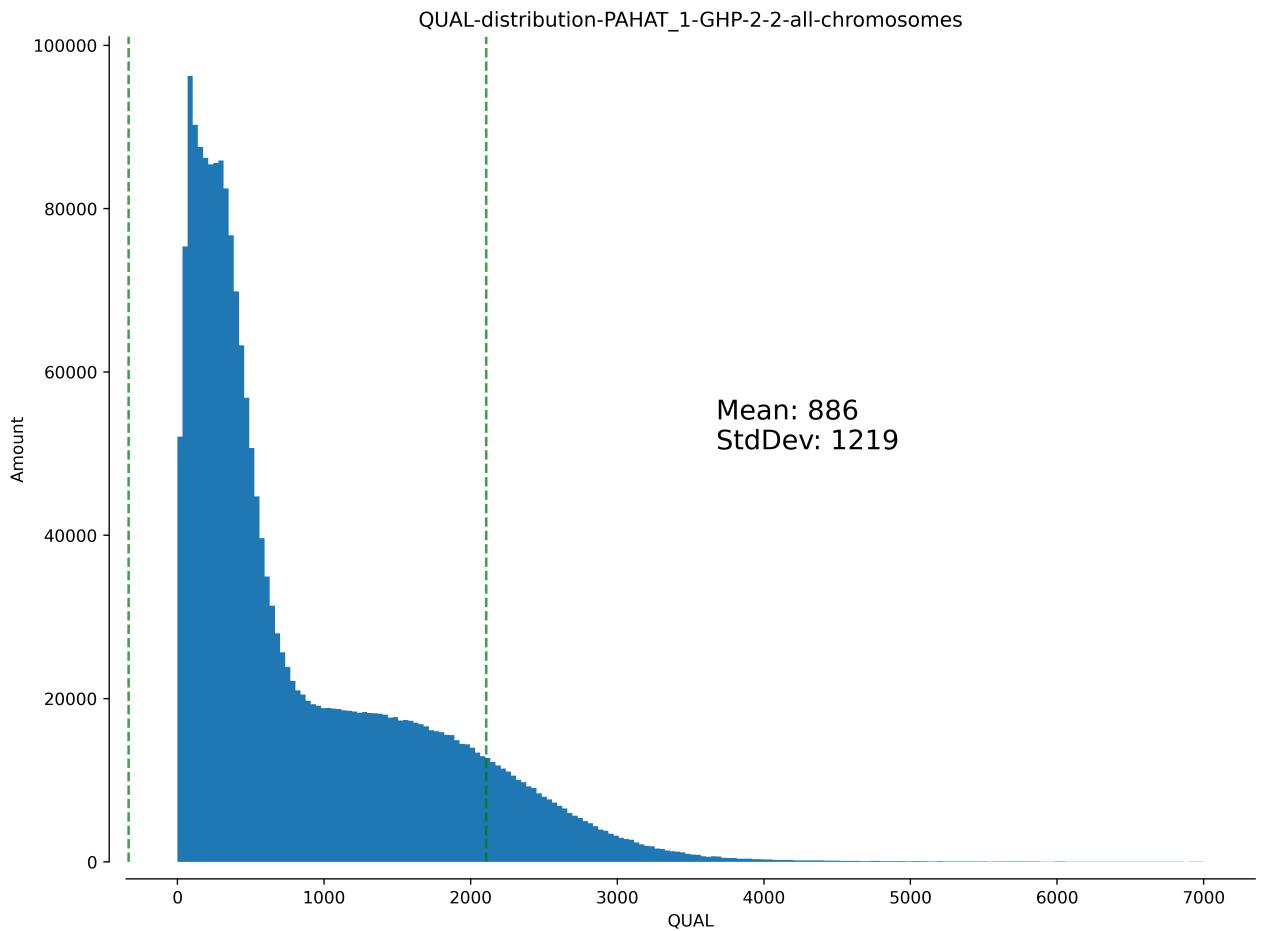
Histograms - DP , QUAL , TYPE and GT Attributes - All Chromosomes - Unfiltered

```
In [10]: plot_variant_hist(samples, vcf_df, 'all', 'DP', bins=200, MSTD=True, xmax=600)
```

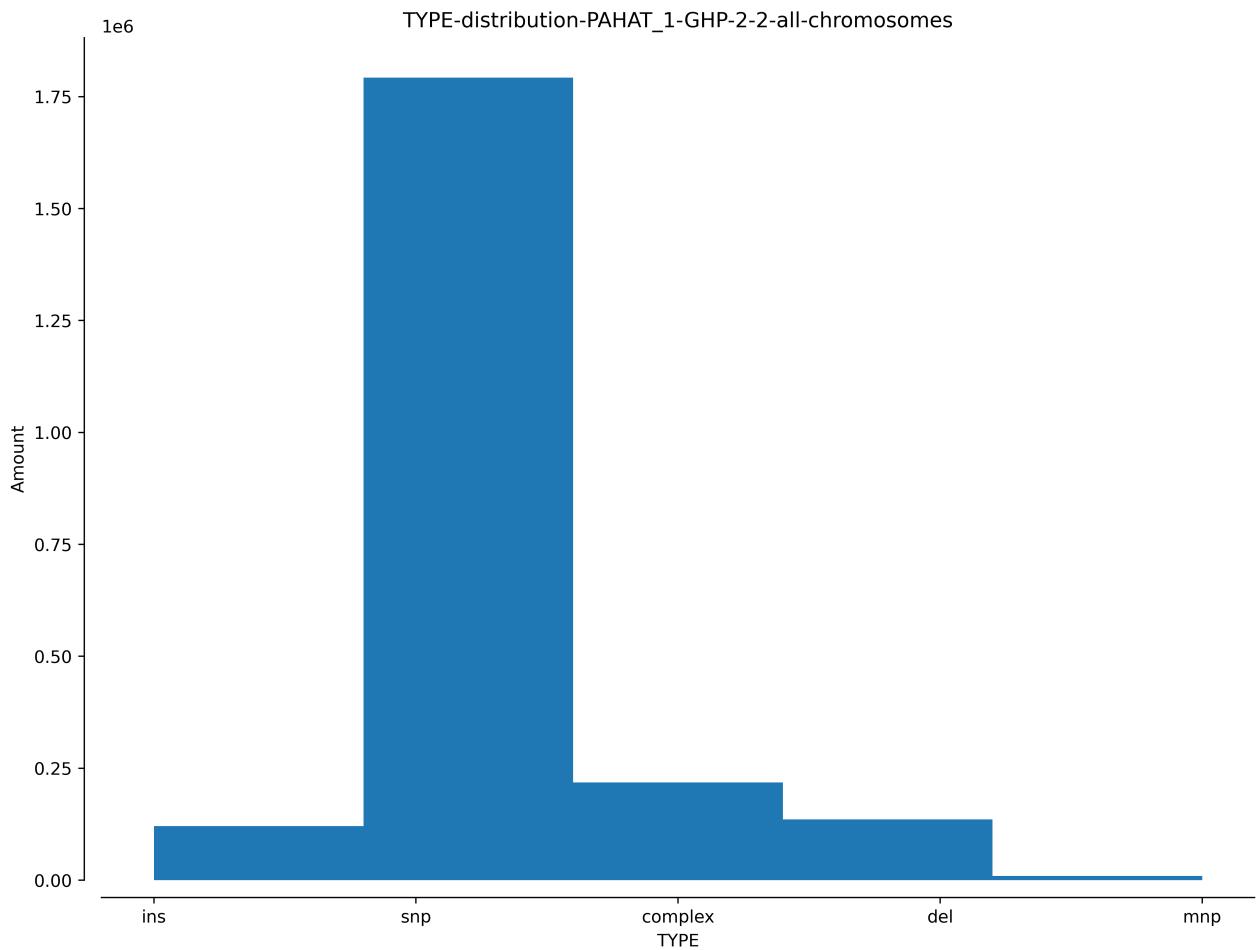


In [11]:

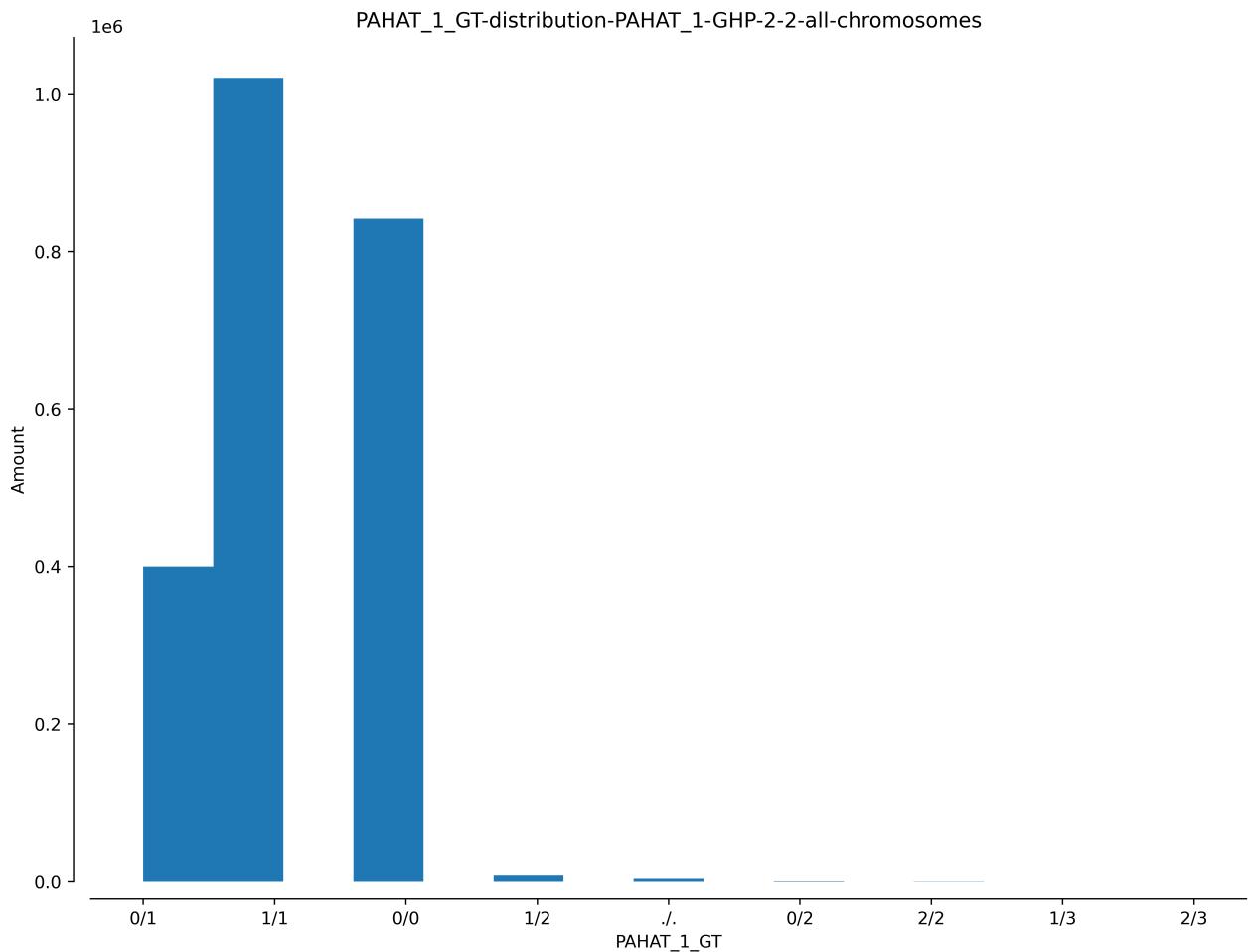
```
plot_variant_hist(samples, vcf_df, 'all', 'QUAL', bins=200, MSTD=True, xmax=700)
```



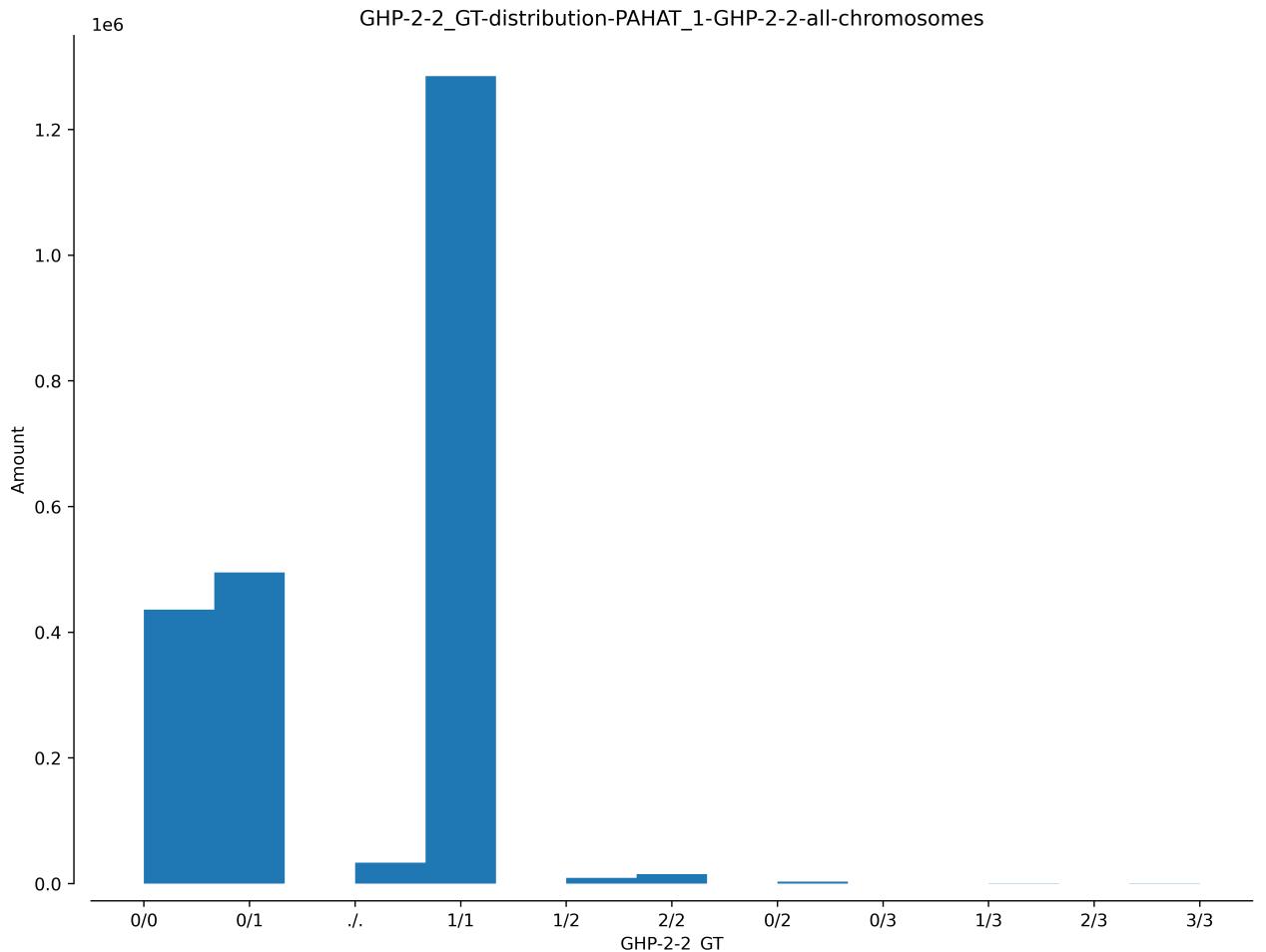
```
In [12]: plot_variant_hist(samples, vcf_df, 'all', 'TYPE', bins=5)
```



```
In [13]: plot_variant_hist(samples, vcf_df, 'all', 'PAHAT_1_GT', bins=15)
```



```
In [14]: plot_variant_hist(samples, vcf_df, 'all', 'GHP-2-2_GT', bins=15)
```



Stacked Bar Plots - RAW

```
In [15]: ct_guide()
```

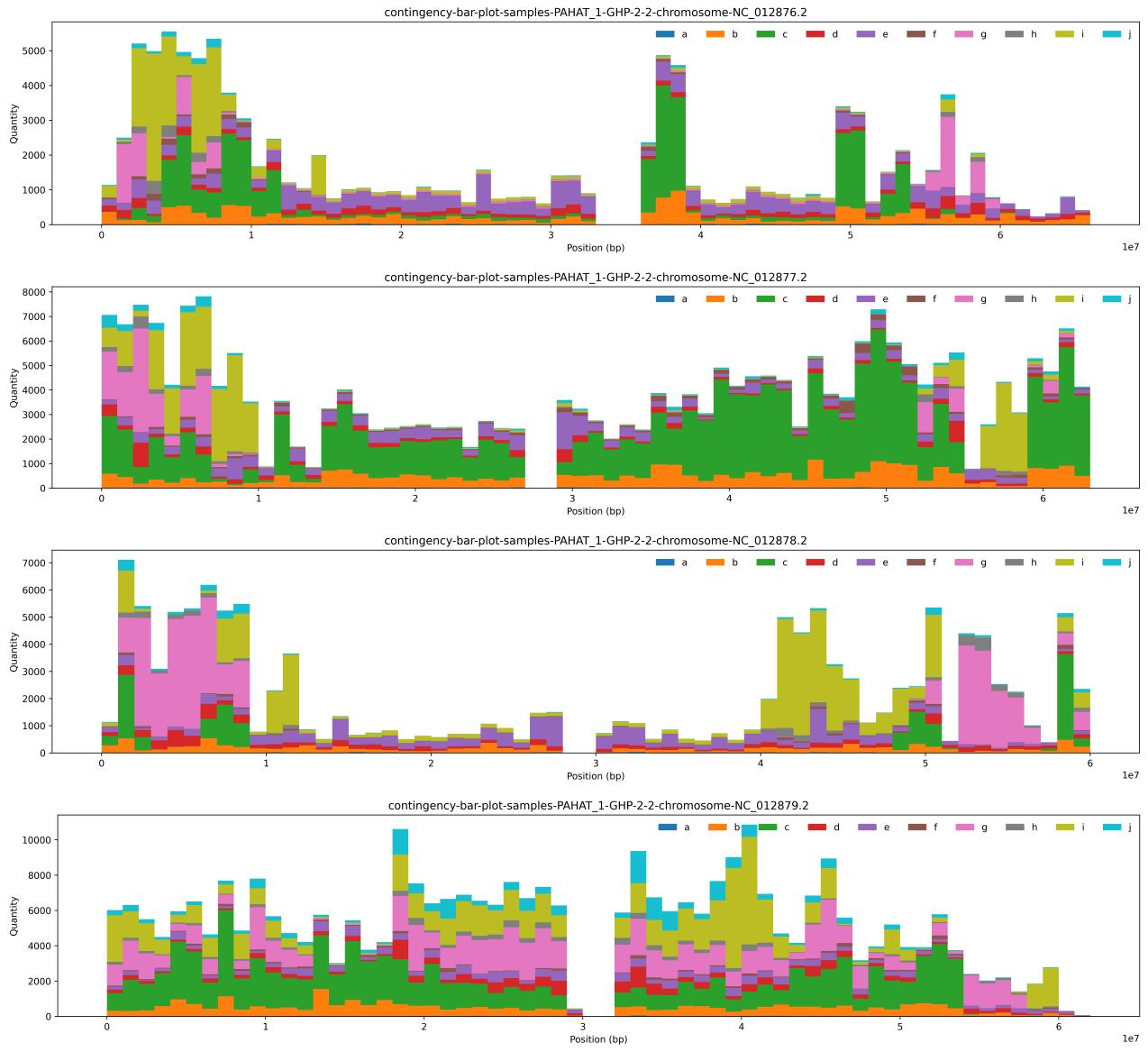
Out[15]:

		Mutant			
		0/0	0/1	1/1	other
Progenitor	0/0	a	b	c	
	0/1	d	e	f	
	1/1	g	h	i	
	other			j	

In [16]:

```
plt.close('all')
window_size = 1000000
CTbarPlots(samples, vcf_df, chrom_len, window_size)
```





PART 1: Filter Out Mitochondria and Chloroplast Chromosomes

Drop Mitochondria and Chloroplast Chromosomes from `vcf_df` and `chrom_len`

```
In [17]: drop_mito_chloro = "CHROM != NC_008360.1, CHROM != NC_008602.1"
vcf_df_00 = filter_vcf(vcf_df, drop_mito_chloro)
vcf_df_00
```

Out[17] :

	CHROM	POS	REF	ALT	QUAL	DP	GHP-2-2_DP	PAI
0	NC_012870.2	91	C	A	113.435997	13	-1	
1	NC_012870.2	111	C	A	226.643997	16	-1	
2	NC_012870.2	744	C	A	178.729996	43	8	
Processing math: 100%	NC_012870.2	871	C	T	499.332001	61	18	

	CHROM	POS	REF	ALT	QUAL	DP	GHP- 2- 2_DP	PAI
4	NC_012870.2	1143	T	G	176.011002	16	9	
...
2276576	NC_012879.2	61191248	GCCA	ACCG	27.908600	71	16	
2276577	NC_012879.2	61191286	C	T	13.749700	70	19	
2276578	NC_012879.2	61191293	G	C	23.474300	65	17	
2276579	NC_012879.2	61233448	GT TAGGGTTAAGGGT	GT TTAGGGTT	432.449005	139	15	
2276580	NC_012879.2	61233618	GT TTAG	GT TTAG	304.154999	117	24	

2276581 rows × 17 columns



In [18]:

```
mito_chloro = ['NC_008360.1', 'NC_008602.1']
chrom_len_00 = chrom_len.drop(mito_chloro)
chrom_len_00
```

Out[18]:

CHROM	LEN
NC_012870.2	80884392
NC_012871.2	77742459
NC_012872.2	74386277
NC_012873.2	68658214
NC_012874.2	71854669
NC_012875.2	61277060
NC_012876.2	65505356
NC_012877.2	62686529
NC_012878.2	59416394
NC_012879.2	61233695

Create Mitochondria and Chloroplast Variants and Chromosome Length Dataframes

In [19]:

```
drop_chrom = "CHROM!=NC_012870.2, CHROM!=NC_012871.2, CHROM!=NC_012872.2, CHROM!=NC_012873.2, CHROM!=NC_012874.2, CHROM!=NC_012875.2, CHROM!=NC_012876.2, CHROM!=NC_012877.2, CHROM!=NC_012878.2, CHROM!=NC_012879.2"
vcf_df_mito_chloro = filter_vcf(vcf_df, drop_chrom)
vcf_df_mito_chloro
```

Out[19]:

	CHROM	POS	REF	ALT	QUAL	DP	GHP- 2- 2_DP	PAHAT_1_DI
0	NC_008360.1	3598	GA AAAAAA ACT	GA AAAAAA ACT	2072.229980	1771	187	85
Processing math: 100%	NC_008360.1	3788	AT	AG	41349.500000	1477	189	69

	CHROM	POS	REF	ALT	QUAL	DP	GHP-2-2_DP	PAHAT_1_DI
2	NC_008360.1	4068	T	A	38302.101562	1387	177	63
3	NC_008360.1	7055	G	A	7958.569824	351	70	16
4	NC_008360.1	9183	G	A	246.141006	26	11	1
...
326	NC_008602.1	111981	CACCCGCAG	CACCTGCAG	294407.000000	13776	4663	461
327	NC_008602.1	115489	AAG	AAA	74523.203125	3835	1330	111
328	NC_008602.1	115682	CA	CC	178725.000000	8256	2511	281
329	NC_008602.1	118002	GGCTGCTACC	GGCAGCTACC	263337.000000	12914	4653	417
330	NC_008602.1	130073	AGAGGT	ACGAGAT	113.835999	12	-1	1

331 rows × 17 columns



```
In [20]: mito_chloro_len = chrom_len.loc[mito_chloro]
mito_chloro_len
```

```
Out[20]:      LEN
              CHROM
NC_008360.1  468628
NC_008602.1  140754
```

Contingency Table - No Mitochondria/Chloroplast

```
In [21]: contingency_table_1 = contingency_table(samples, vcf_df_00, 'all')
```

Contingency Table - Chromosome all

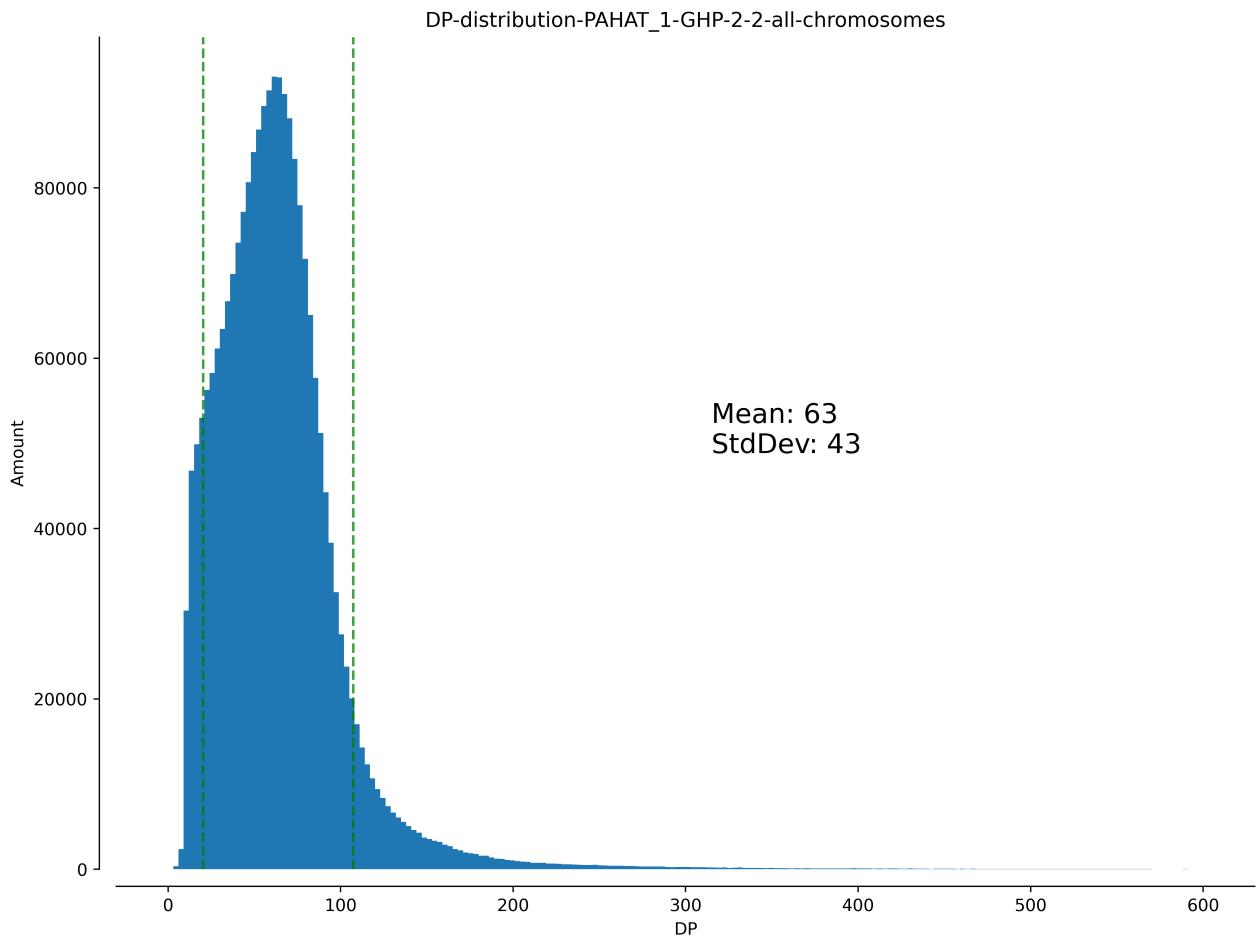
PAHAT_1_GT	GHP-2-2_GT			
	0/0	0/1	1/1	other
0/0	7585	212887	620457	68924
0/1	119369	225903	42168	68924
1/1	307108	54103	618077	68924
other	68924	68924	68924	68924

GT Plot - No Mitochondria/Chloroplast

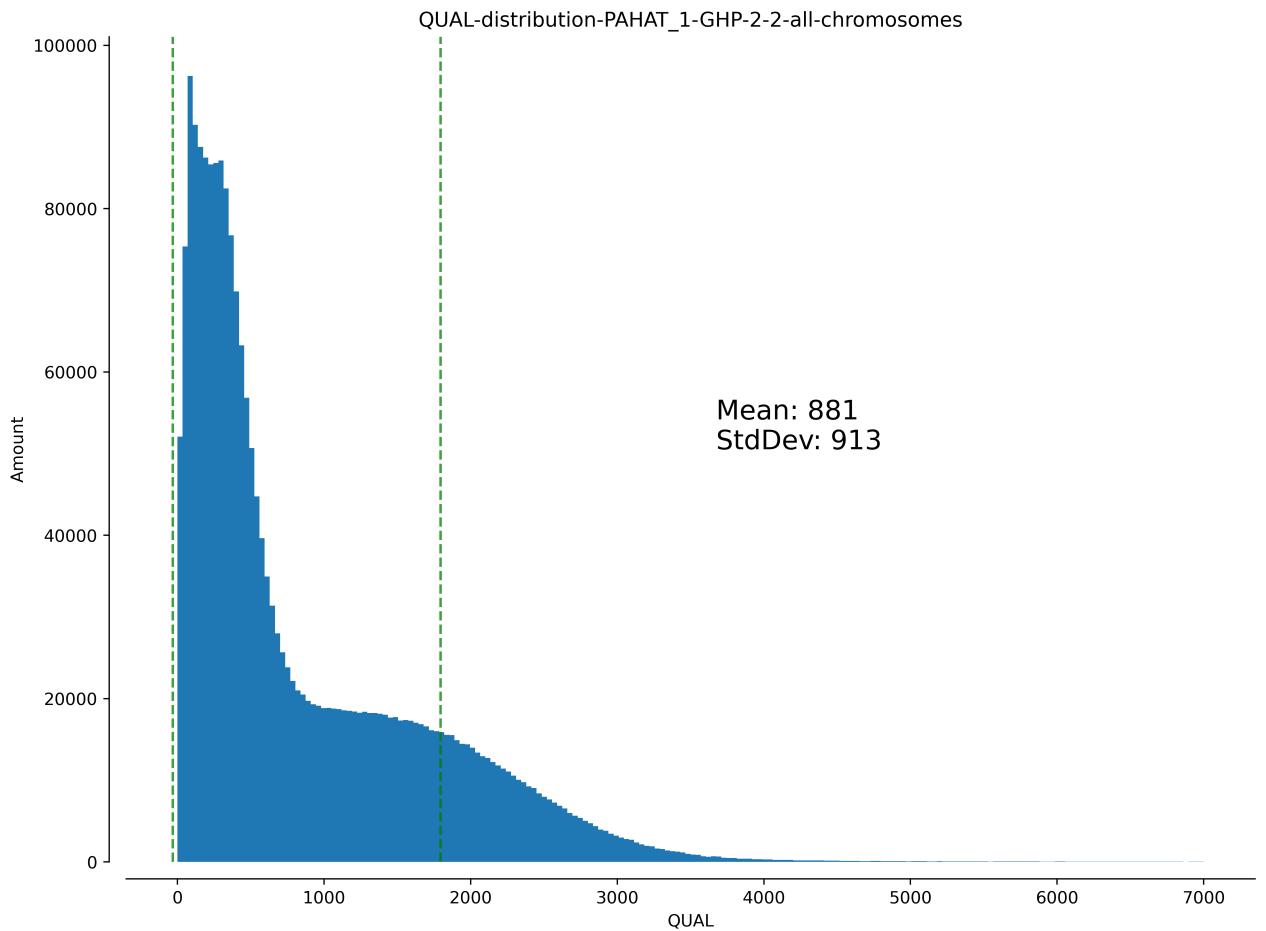
```
In [22]: # plt.close('all')
# GTplot(samples, vcf_df_00, chrom_len_00)
```

Histograms - DP , QUAL , TYPE and GT Attributes - No Mitochondria/Chloroplast

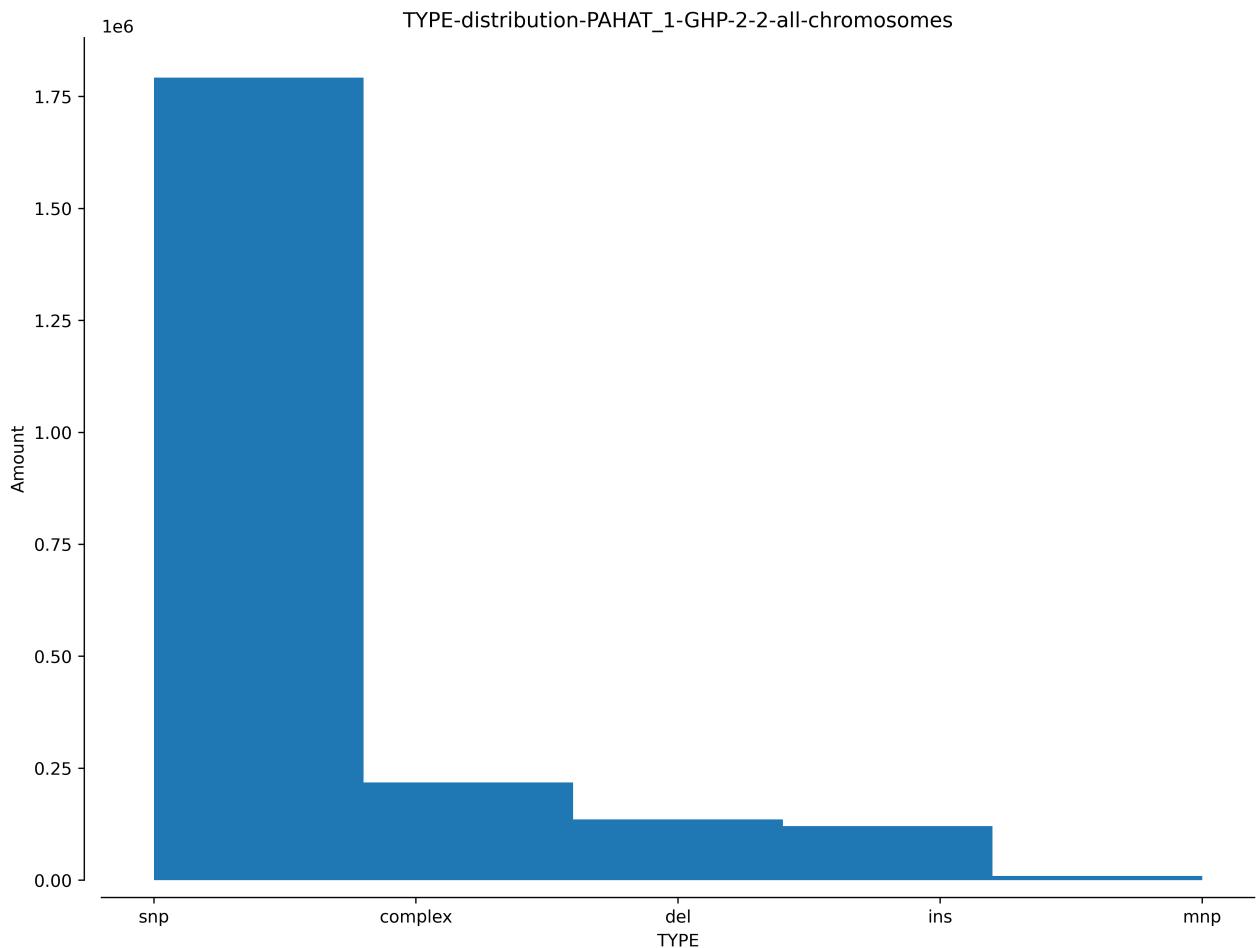
```
In [23]: plot.variant_hist(samples, vcf_df_00, 'all', 'DP', bins=200, MSTD=True, xmax=600)
Processing math: 100%
```



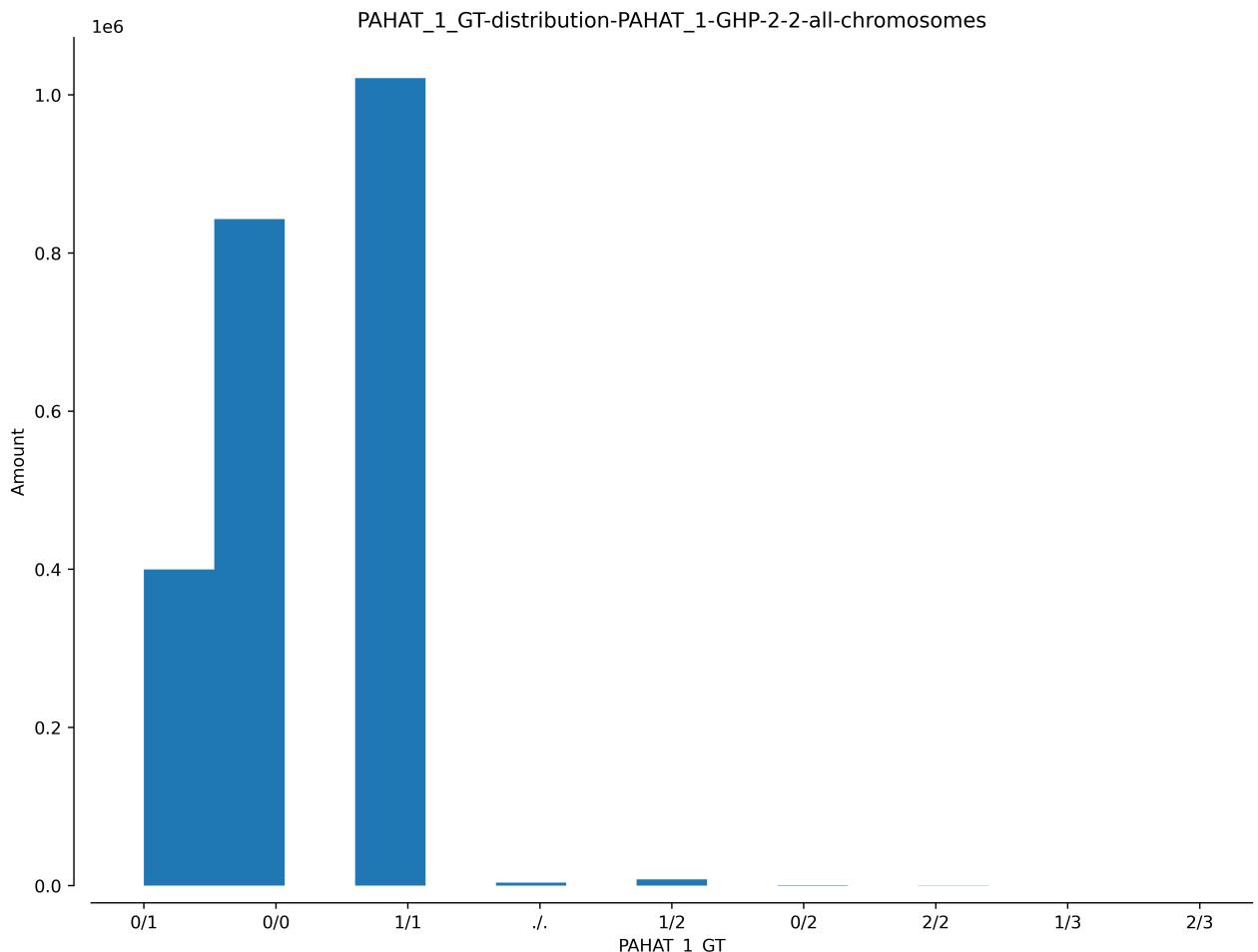
In [24]: `plot_variant_hist(samples, vcf_df_00, 'all', 'QUAL', bins=200, MSTD=True, xmax=`



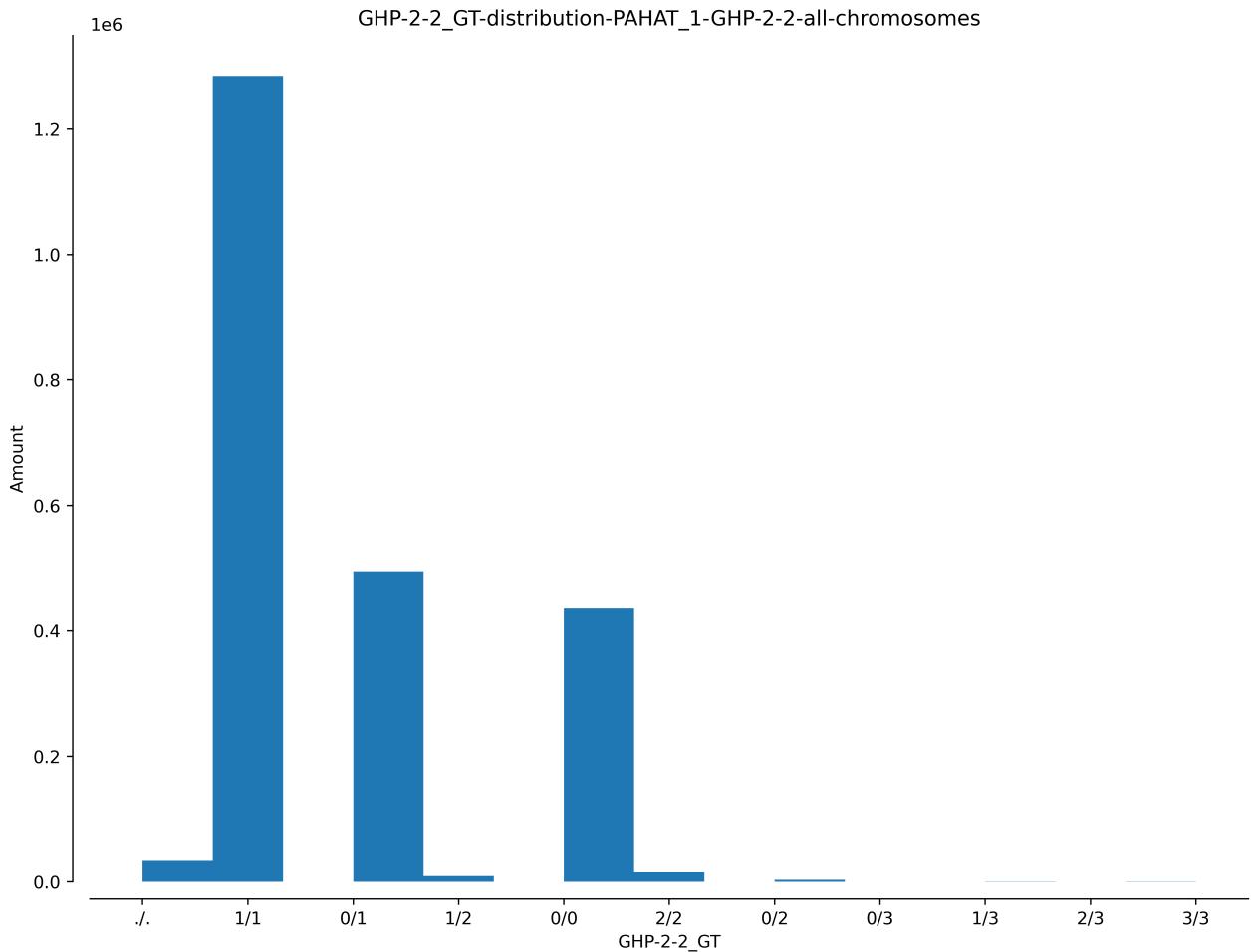
```
In [25]: plot_variant_hist(samples, vcf_df_00, 'all', 'TYPE', bins=5)
```



```
In [26]: plot_variant_hist(samples, vcf_df_00, 'all', 'PAHAT_1_GT', bins=15)
```



```
In [27]: plot_variant_hist(samples, vcf_df_00, 'all', 'GHP-2-2_GT', bins=15)
```



PART 2: Cutting Off by Mean±2StdDev Histograms of *DP* Attribute

In [28]:

```
cutoff_left = vcf_df_00.DP.mean() - (2 * vcf_df_00.DP.std())
cutoff_right = vcf_df_00.DP.mean() + (2 * vcf_df_00.DP.std())

filter_dp = "DP >= %i, DP <= %i" % (cutoff_left, cutoff_right)
print(filter_dp)

vcf_df_01 = filter_vcf(vcf_df_00, filter_dp)
vcf_df_01
```

DP >= -23, DP <= 150

Out[28]:

	CHROM	POS	REF	ALT	QUAL	DP	GHP-2-2_DP	PAI
0	NC_012870.2	91	C	A	113.435997	13	-1	
1	NC_012870.2	111	C	A	226.643997	16	-1	
2	NC_012870.2	744	C	A	178.729996	43	8	
3	NC_012870.2	871	C	T	499.332001	61	18	
4	NC_012870.2	1143		T		G	176.011002	16

Processing math: 100%

...

...

...

...

...

...

...

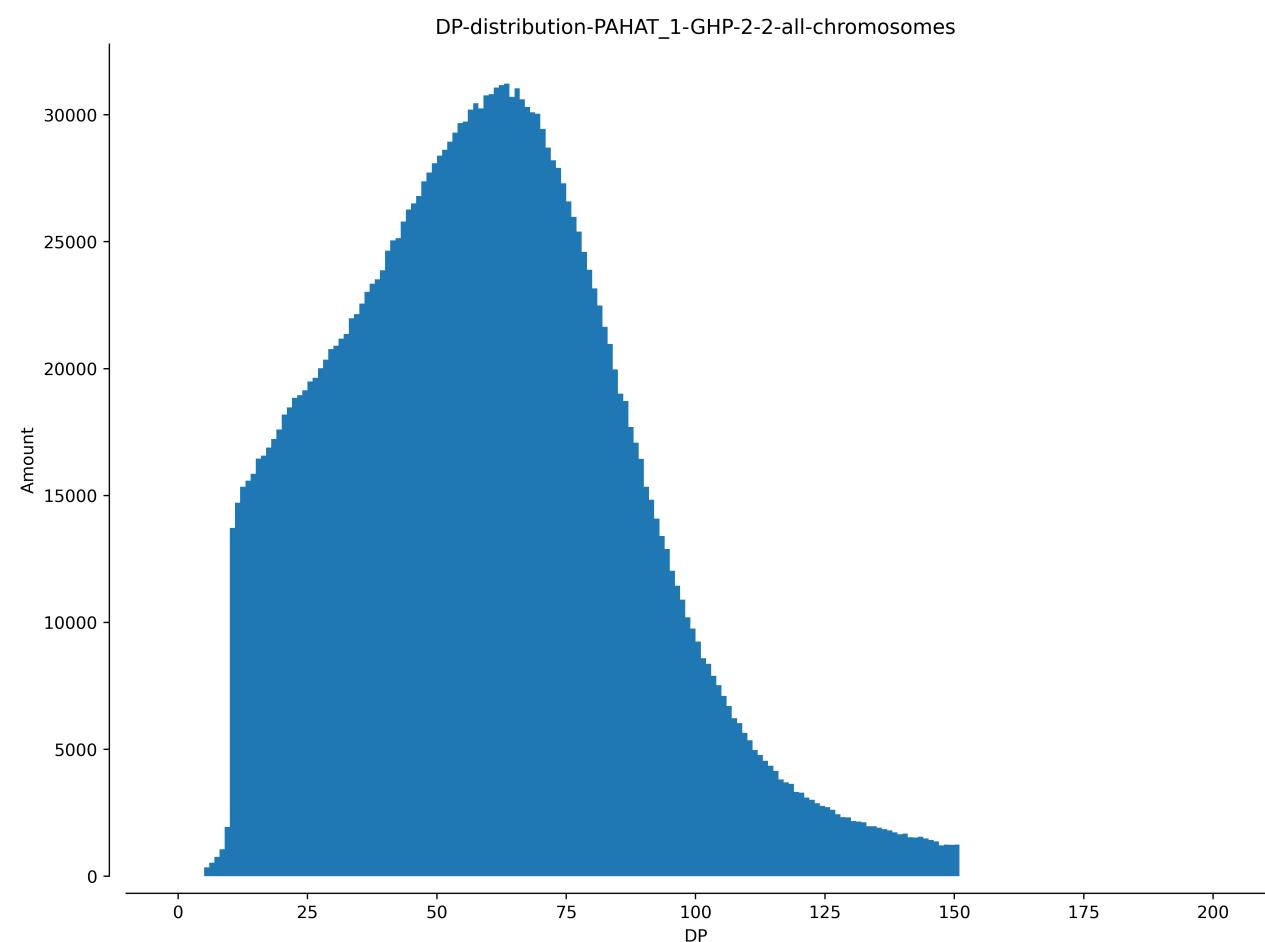
	CHROM	POS	REF	ALT	QUAL	DP	GHP- 2- 2_DP	PAI
2217645	NC_012879.2	61191248	GCCA	ACCG	27.908600	71	16	
2217646	NC_012879.2	61191286		C	T	13.749700	70	19
2217647	NC_012879.2	61191293		G	C	23.474300	65	17
2217648	NC_012879.2	61233448	GTTCAGGGTTAAGGGT	GTTCAGGGTT	432.449005	139	15	
2217649	NC_012879.2	61233618		GTTTAG	GTTTAG	304.154999	117	24

2217650 rows × 17 columns

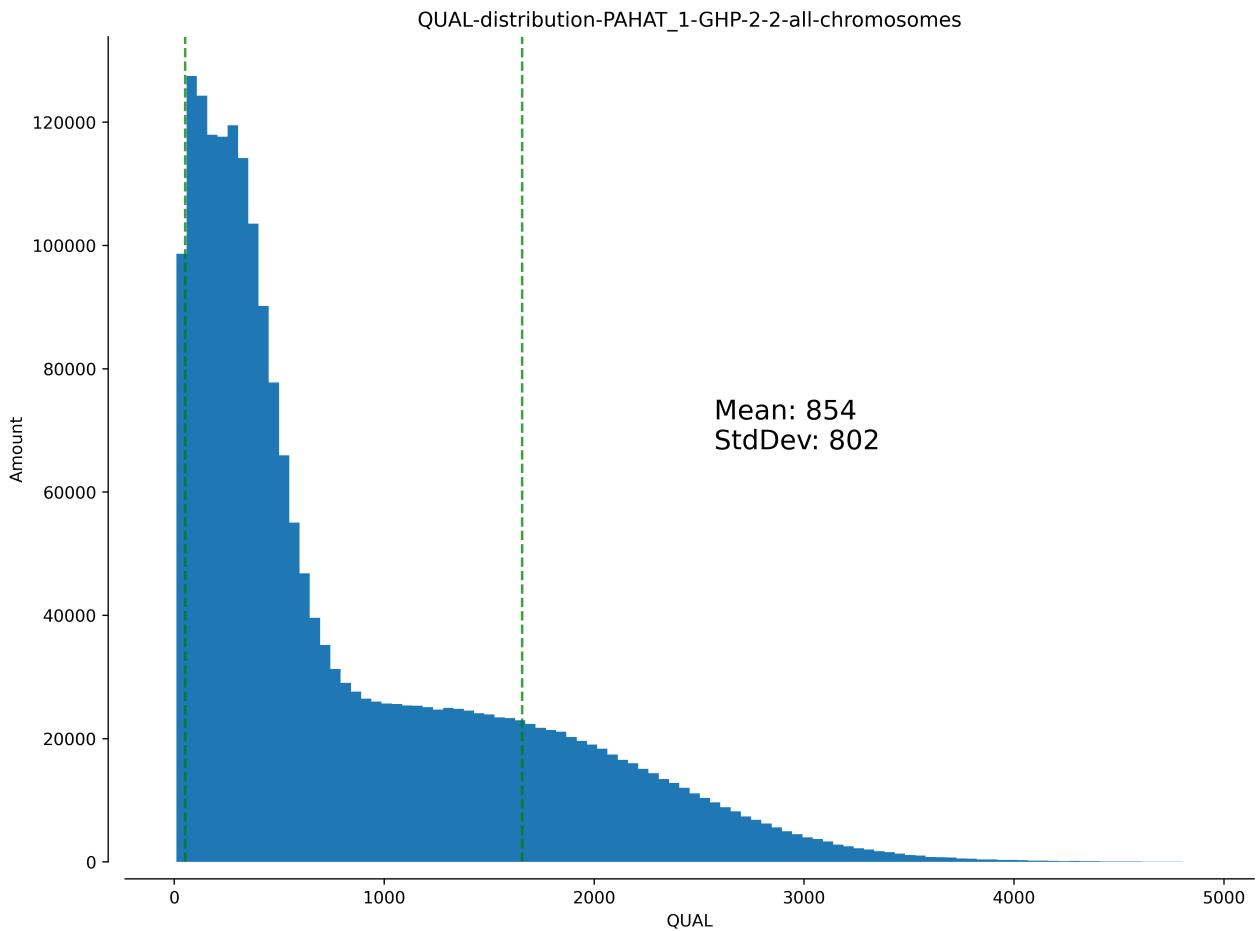


Verify DP Histogram Cutoff Off by Mean±StdDev

In [29]: `plot_variant_hist(samples, vcf_df_01, 'all', 'DP', bins=200, xmax=200)`



In [30]: `plot_variant_hist(samples, vcf_df_01, 'all', 'QUAL', bins=100, MSTD=True)`



Contingency Table After DP Cutoff by Mean±StdDev

```
In [31]: contingency_table_2 = contingency_table(samples, vcf_df_01, 'all')
```

Contingency Table - Chromosome all

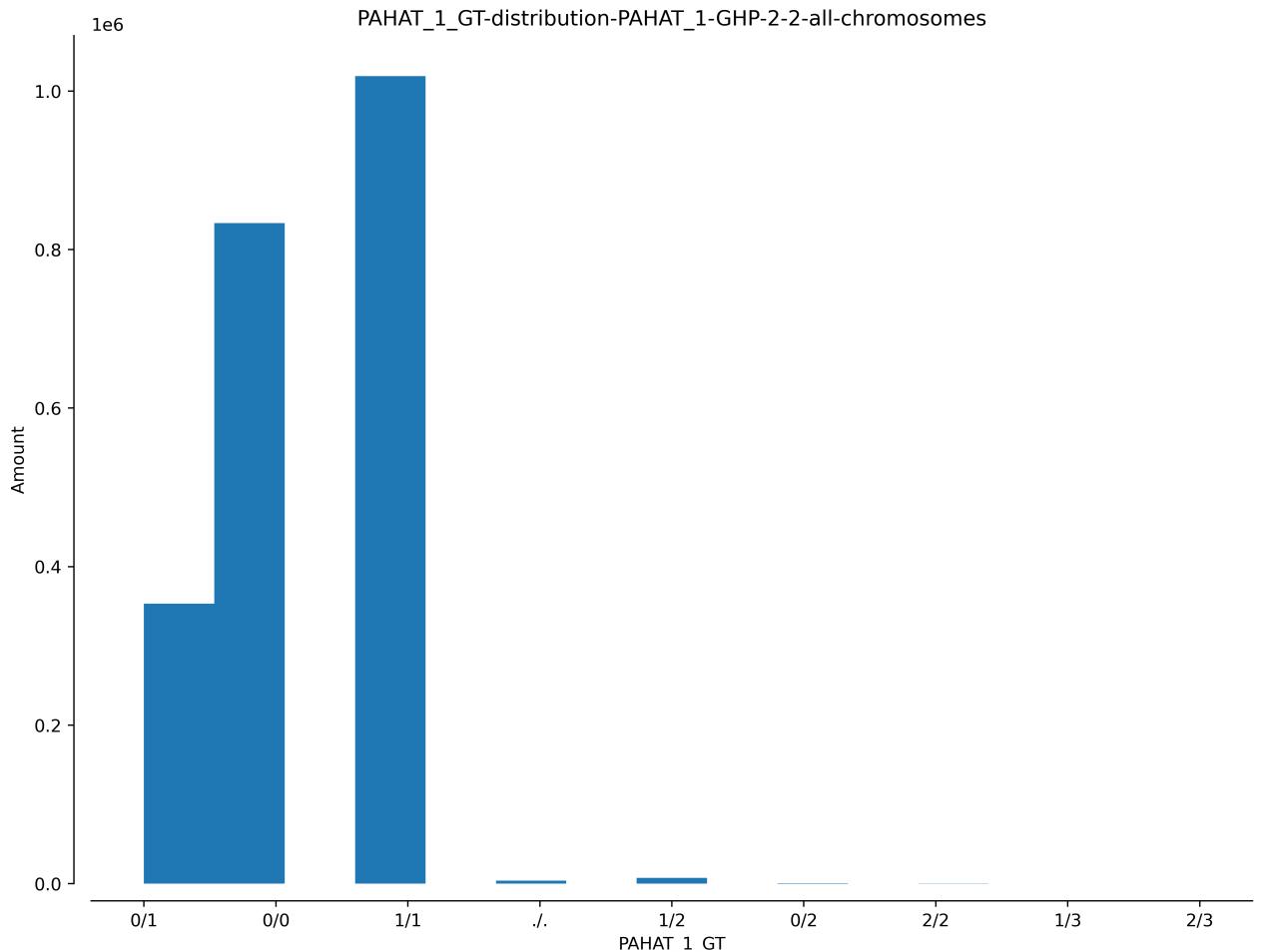
		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	7349	206016	618142	67287
	0/1	110174	191367	40272	67287
	1/1	306772	53835	616436	67287
	other	67287	67287	67287	67287

GT Plot After DP Cutoff by Mean±StdDev

```
In [32]: # plt.close('all')
# GTplot(samples, vcf_df_01, chrom_len_00)
```

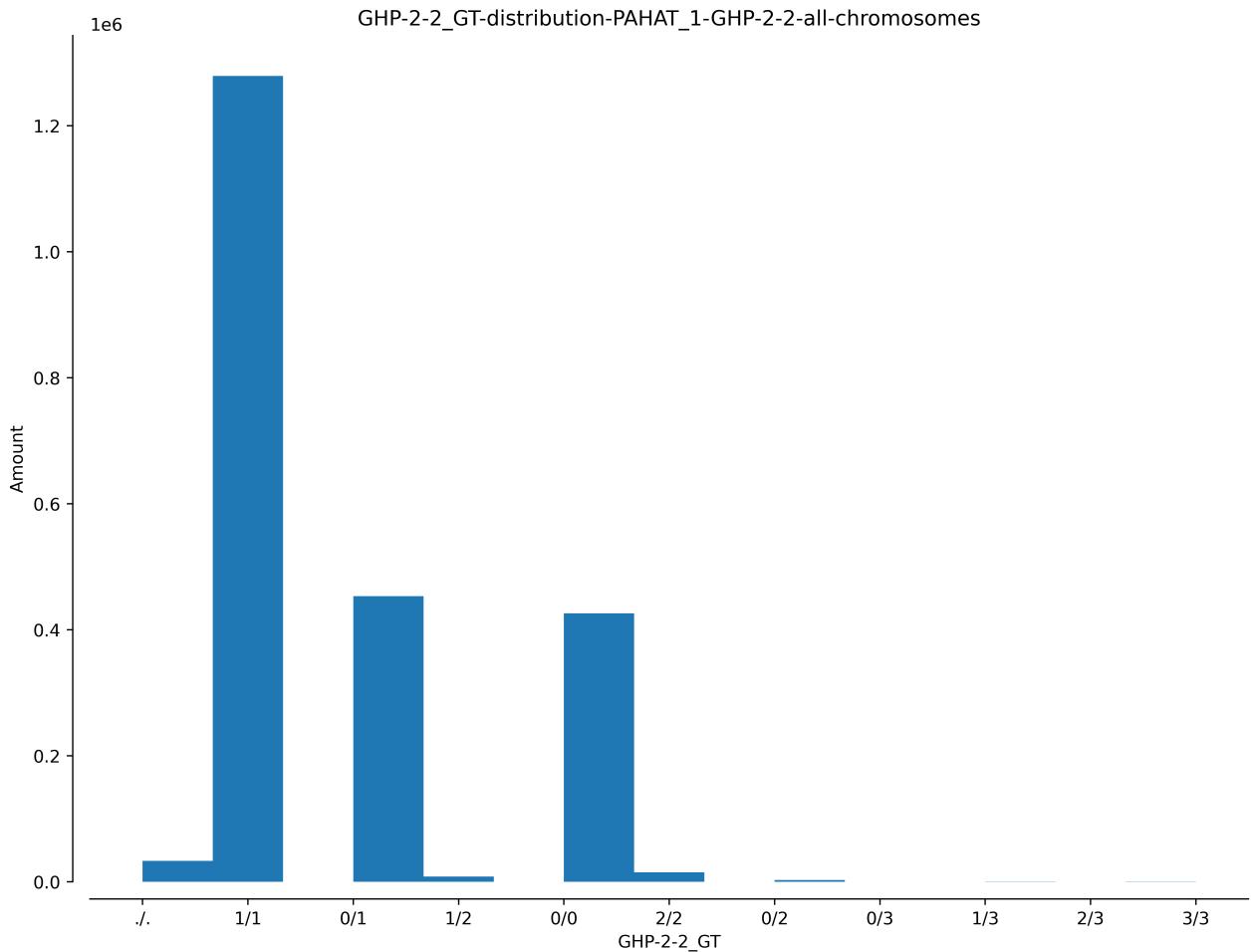
Histogram 'GT' Attribute after DP Cutoff

```
In [33]: plot_variant_hist(samples, vcf_df_01, 'all', 'PAHAT_1_GT', bins=15)
```



In [34]:

```
plot_variant_hist(samples, vcf_df_01, 'all', 'GHP-2-2_GT', bins=15)
```



PART 3: Extract $\in s$ and ∂ from TYPE Attribute

Extract ins and del TYPE Attribute

In [35]:

```
extract_type = "TYPE != snp, TYPE != complex, TYPE != mnp"
vcf_df_02 = filter_vcf(vcf_df_01, extract_type)
vcf_df_02
```

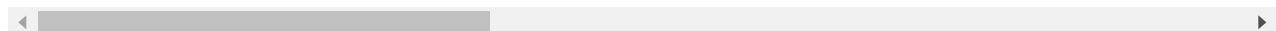
Out[35]:

	CHROM	POS	REF
0	NC_012870.2	1639	ACCCCCCCCAGCA
1	NC_012870.2	2881	GTTTTTTTGTCTG
2	NC_012870.2	3806	ATTTTTTTTACA
3	NC_012870.2	4512	TATTTCCCACACAAACCACACTCACCCGCTAAAGTGAGCTACAGAT...
4	NC_012870.2	10162	TGGGGGCA
...
254720	NC_012879.2	60955570	TAAAAAAAAAAAAAGAAAAGAAAGGGGGAGT TAA/
254721	NC_012879.2	60989473	GATACCCAAA
	NC_012879.2	61137974	ATTTTTTTAGA

Processing math: 100%

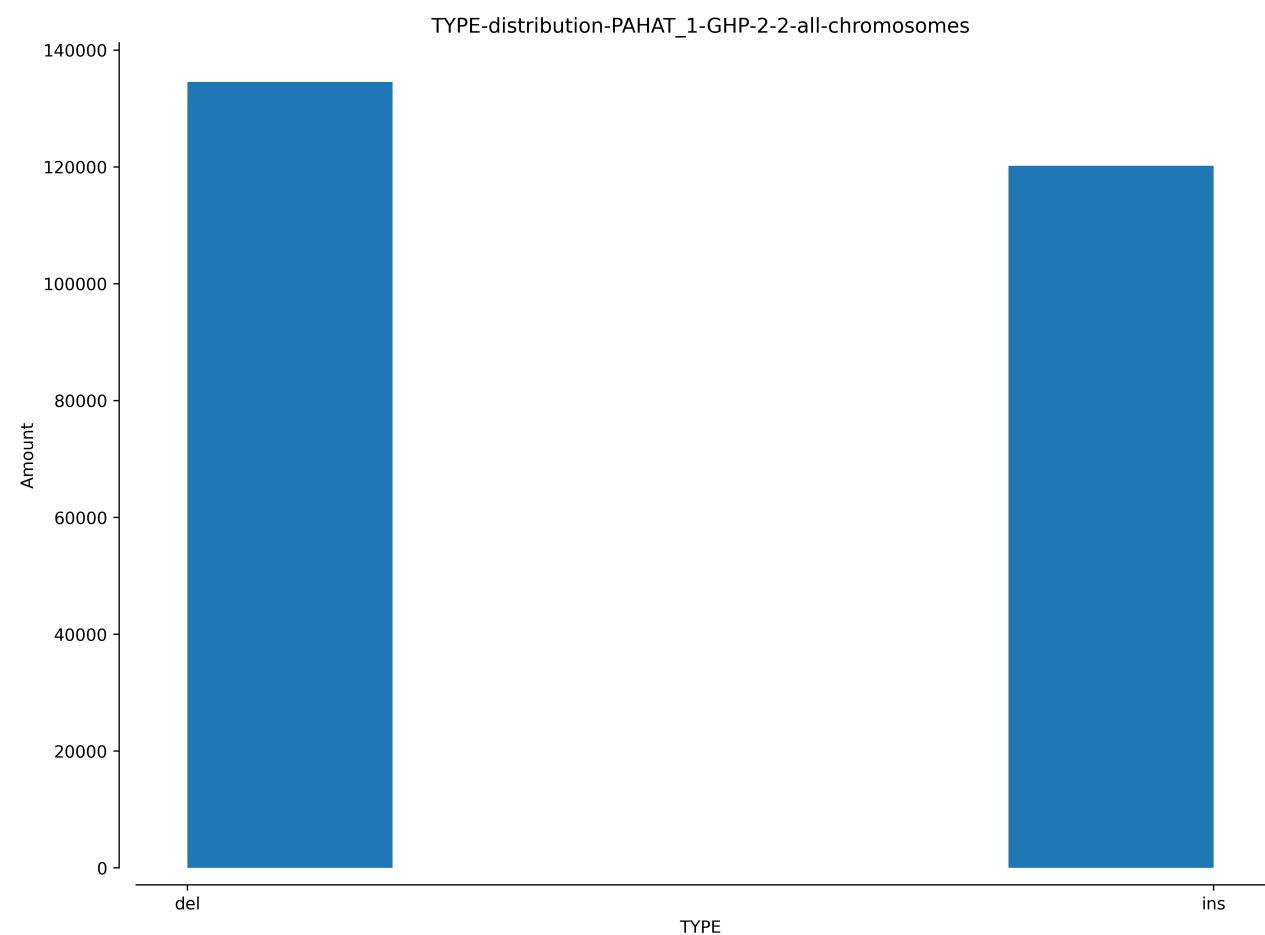
	CHROM	POS	REF
254723	NC_012879.2	61160465	GCG
254724	NC_012879.2	61233618	GTTTTAG

254725 rows × 17 columns



TYPE ins and del Histogram Verification

In [36]: `plot_variant_hist(samples, vcf_df_02, 'all', 'TYPE', bins=5)`



Contingency Table - ins and del TYPE only

In [37]: `contingency_table_3 = contingency_table(samples, vcf_df_02, 'all')`

Contingency Table - Chromosome all

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	305	15850	69051	13418
	0/1	8133	9738	4646	13418
	1/1	48313	7313	77958	13418
	other	13418	13418	13418	13418

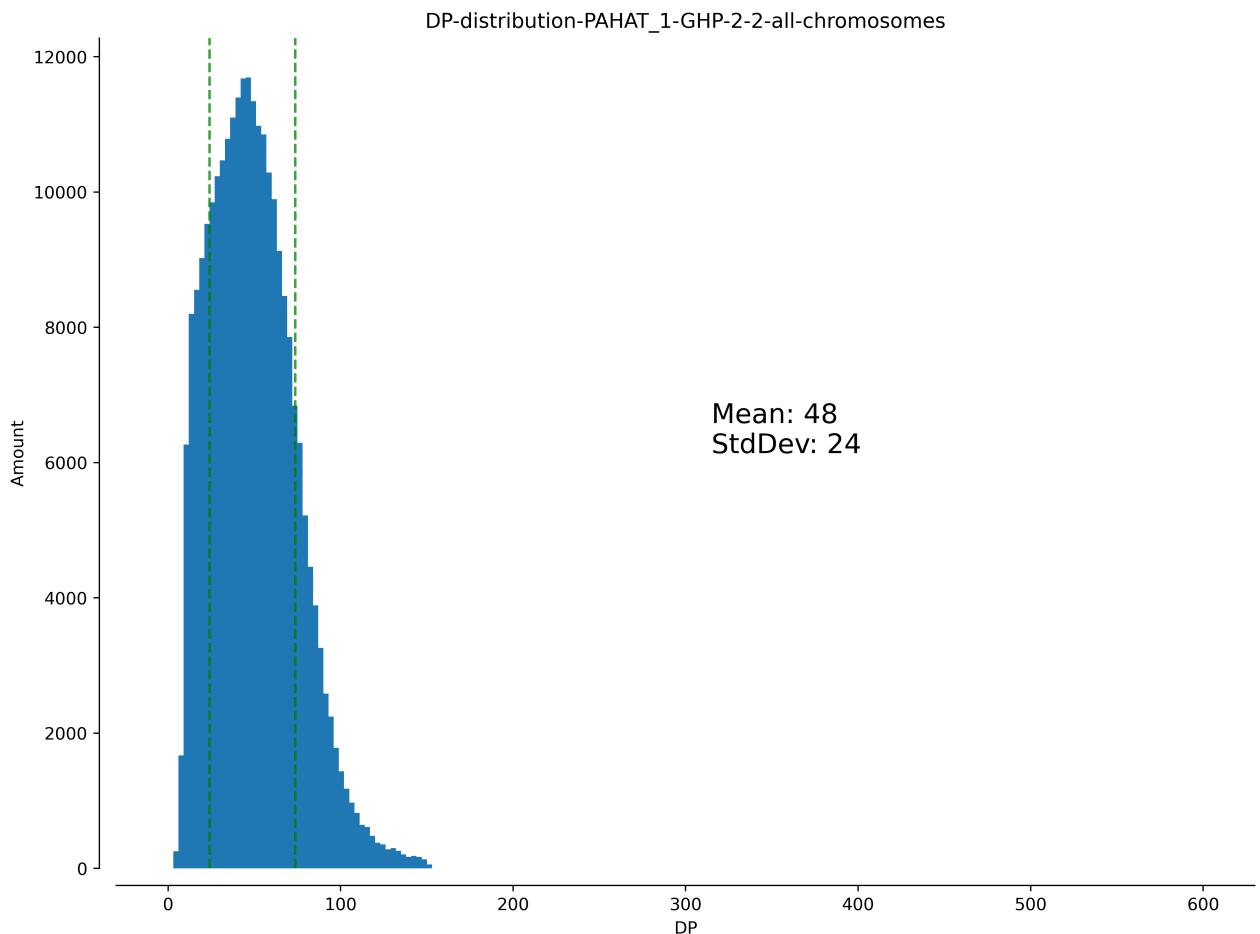
Processing math: 100%

GT Plot - ins and del TYPE only

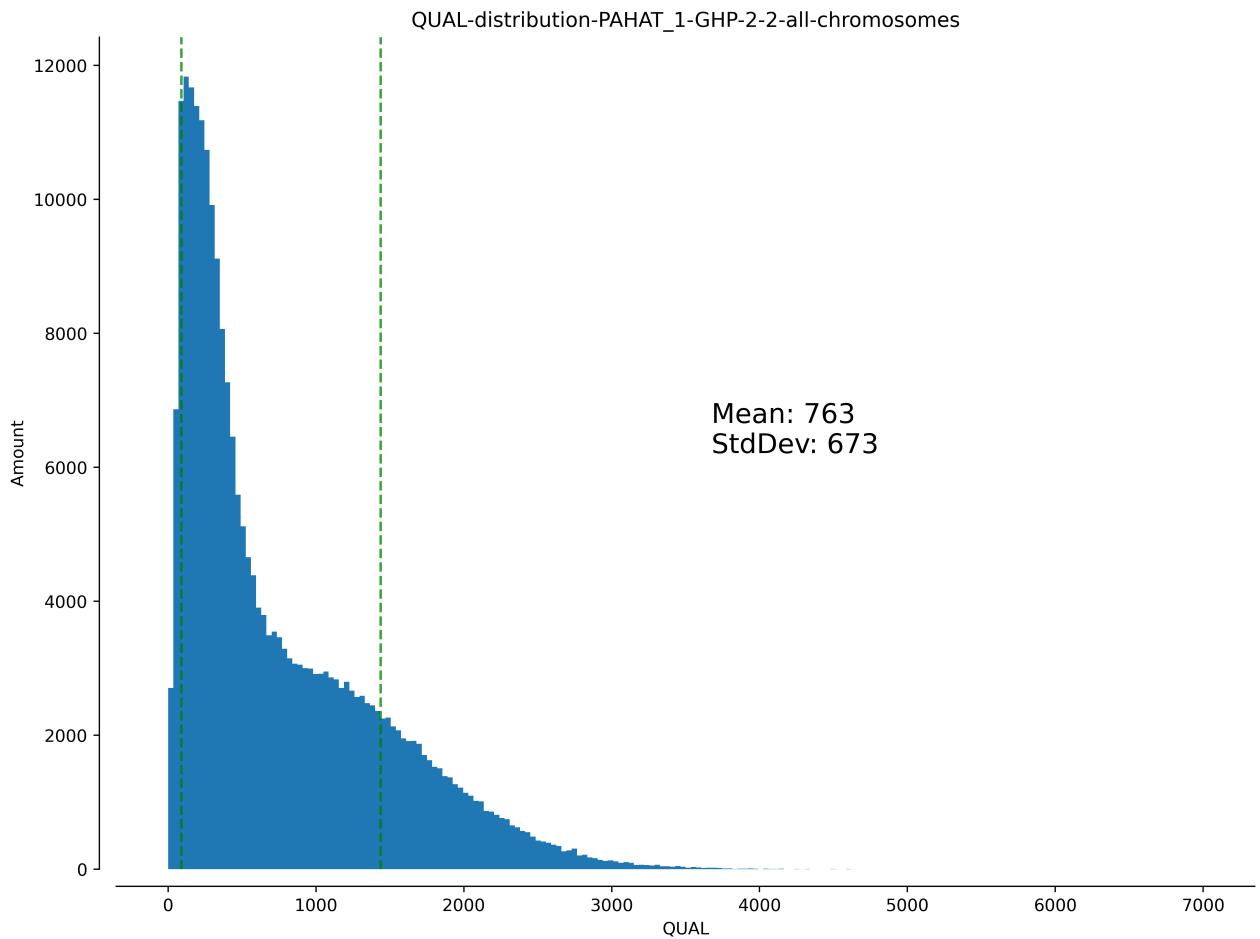
```
In [38]: # plt.close('all')
# GTplot(samples, vcf_df_02, chrom_len_00)
```

Histograms - DP, QUAL, and GT Attributes after TYPE Filtering

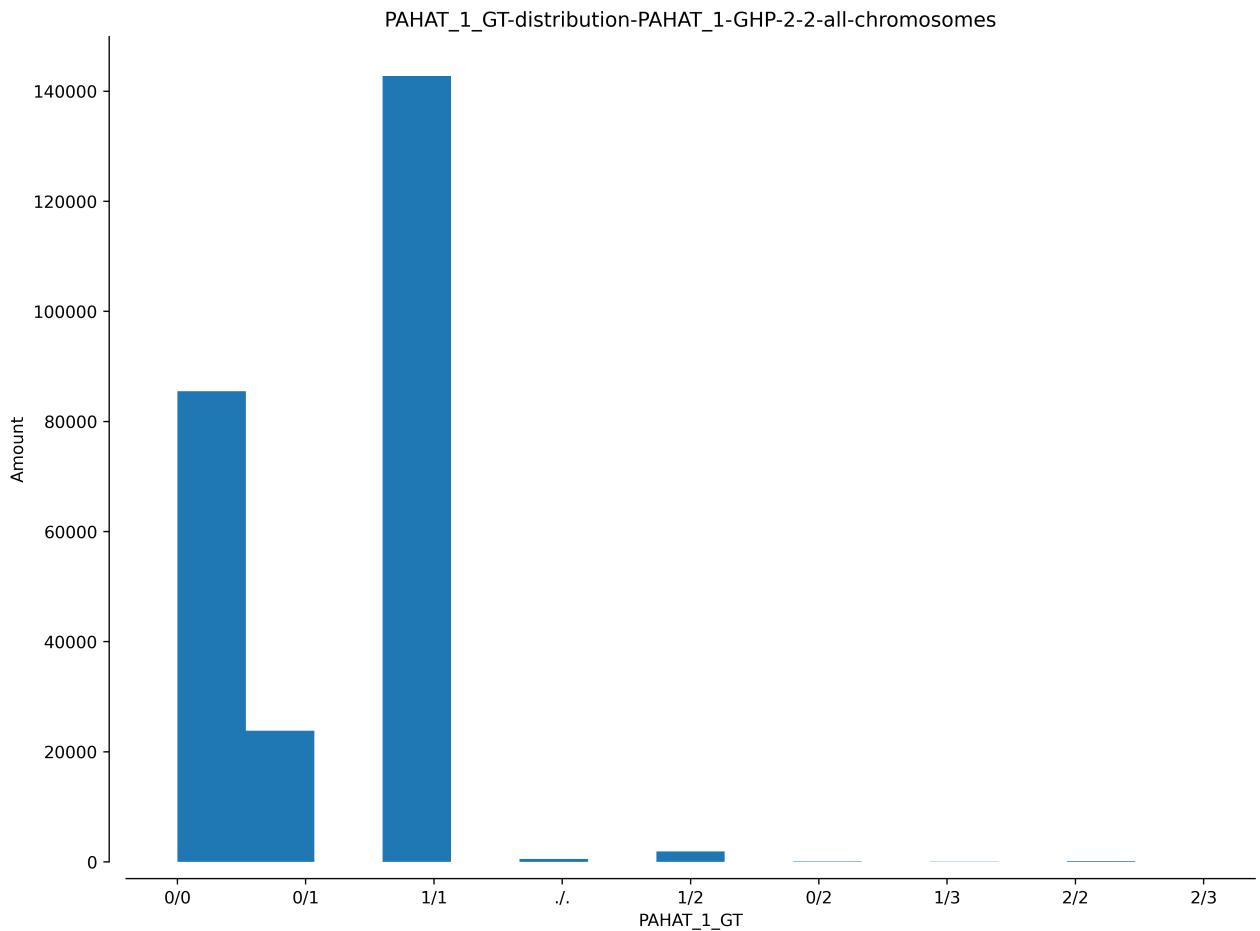
```
In [39]: plot_variant_hist(samples, vcf_df_02, 'all', 'DP', bins=200, MSTD=True, xmax=600)
```



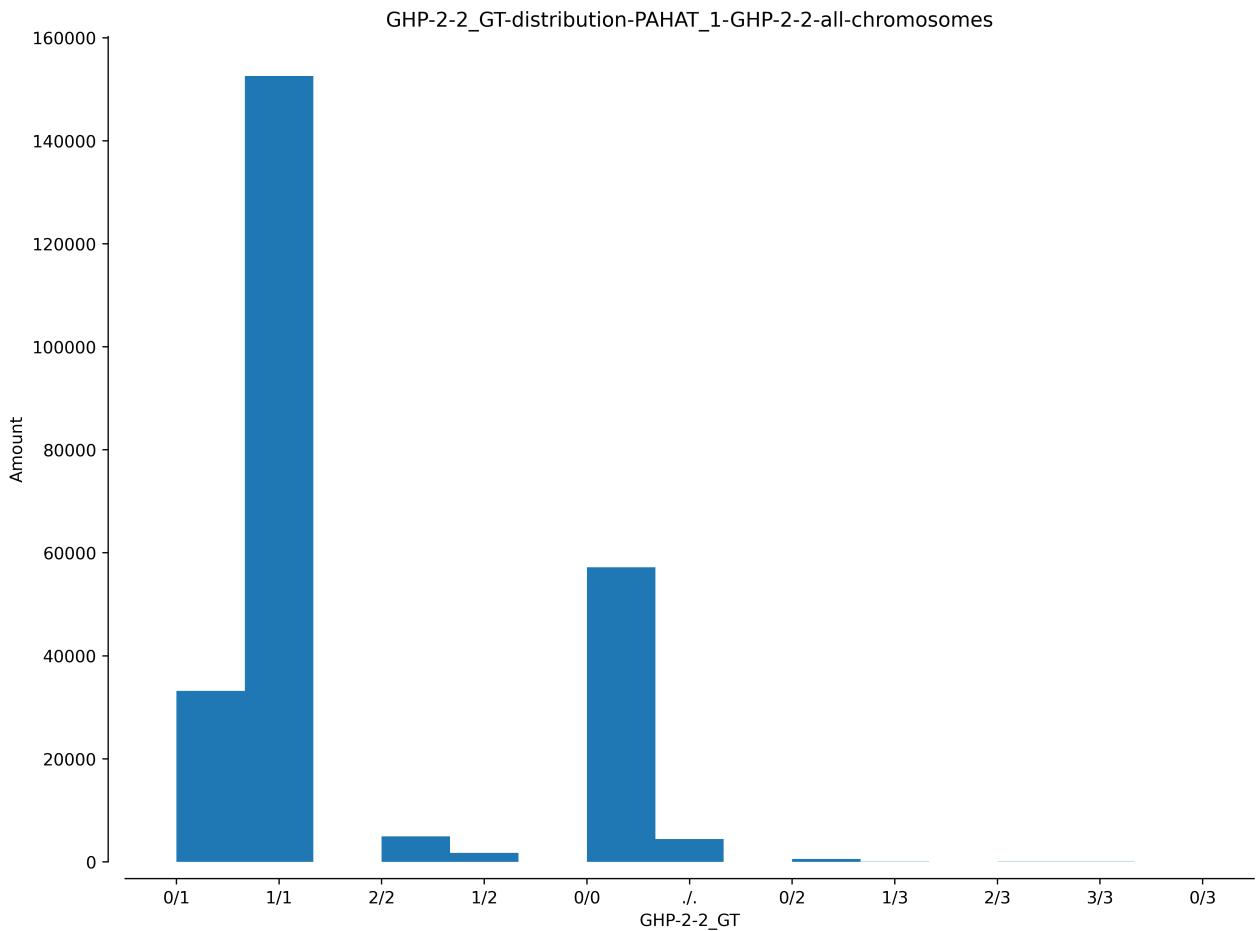
```
In [40]: plot_variant_hist(samples, vcf_df_02, 'all', 'QUAL', bins=200, MSTD=True, xmax=
```



```
In [41]: plot_variant_hist(samples, vcf_df_02, 'all', 'PAHAT_1_GT', bins=15)
```



```
In [42]: plot_variant_hist(samples, vcf_df_02, 'all', 'GHP-2-2_GT', bins=15)
```



PART 4: Cutting Off by Mean±StdDev Histograms of QUAL Attribute

In [43]:

```
# cutoff_left = vcf_df_02.QUAL.mean() - vcf_df_02.QUAL.std()
# cutoff_right = vcf_df_02.QUAL.mean() + vcf_df_02.QUAL.std()

# filter_qual = "QUAL >= %i, QUAL <= %i" % (cutoff_left, cutoff_right)
# print(filter_qual)

# vcf_df_03 = filter_vcf(vcf_df_02, filter_qual)
# vcf_df_03
```

Verify DP and QUAL Histograms after QUAL Cutoff Off by Mean±StdDev

In [44]:

```
# plot_variant_hist(samples, vcf_df_03, 'all', 'DP', bins=200, xmax=200)
```

In [45]:

```
# plot_variant_hist(samples, vcf_df_03, 'all', 'QUAL', bins=100, xmax=3500)
```

Contingency Table After QUAL Cutoff by Mean±StdDev

In [46]:

```
# contingency_table_4 = contingency_table(samples, vcf_df_03, 'all')
# contingency_table_4
```

Processing math: 100%

GT Plot After QUAL Cutoff by Mean±StdDev

```
In [47]: # plt.close('all')
# GTplot(samples, vcf_df_03, chrom_len_00)
```

Histograms after QUAL Cutoff by Mean±StdDev

```
In [48]: # plot_variant_hist(samples, vcf_df_03, 'all', 'PAHAT_1_GT', bins=9)
```

```
In [49]: # plot_variant_hist(samples, vcf_df_03, 'all', 'GHP-2-2_GT', bins=9)
```

PART 5: Filtering GTs 0/0, 1/1, 'Other'

Filter out where samples GTs are the same (0/0, 1/1) and have 'Other'

```
In [50]: PAHAT_gts_filter = "PAHAT_1_GT != ./., PAHAT_1_GT != 0/2, PAHAT_1_GT != 1/2, PAHAT_1_GT != 2/2"
vcf_df_04 = filter_vcf(vcf_df_02, PAHAT_gts_filter)

GHP_gts_filter = "GHP-2-2_GT != ./., GHP-2-2_GT != 0/2, GHP-2-2_GT != 1/2, GHP-2-2_GT != 2/2"
vcf_df_04 = filter_vcf(vcf_df_04, GHP_gts_filter)

genotypes = ['0/0', '1/1']
for genotype in genotypes:
    vcf_df_04 = filter_similar_gt(samples, vcf_df_04, genotype)

vcf_df_04
```

Out[50]:

	CHROM	POS	REF
0	NC_012870.2	1639	ACCCCCCCCAGCA
1	NC_012870.2	2881	GTTTTTTTGTCTG
2	NC_012870.2	3806	ATTTTTTTTACA
3	NC_012870.2	4512	TATTTCACCACACAAACCACACTCACCCGCTAAAGTGAGCTACAGAT...
4	NC_012870.2	10162	TGGGGGCA
...
163039	NC_012879.2	60955570	TAAAAAAAAAAAAGAAAAGAAAGGGGGAGT TAA/
163040	NC_012879.2	60989473	GATACCCAAA
163041	NC_012879.2	61137974	ATTTTTTTAGA
163042	NC_012879.2	61160465	GCG
163043	NC_012879.2	61233618	GTTTTAG

163044 rows × 17 columns

```
In [51]: contingency_table_5 = contingency_table(samples, vcf_df_04, 'all')
```

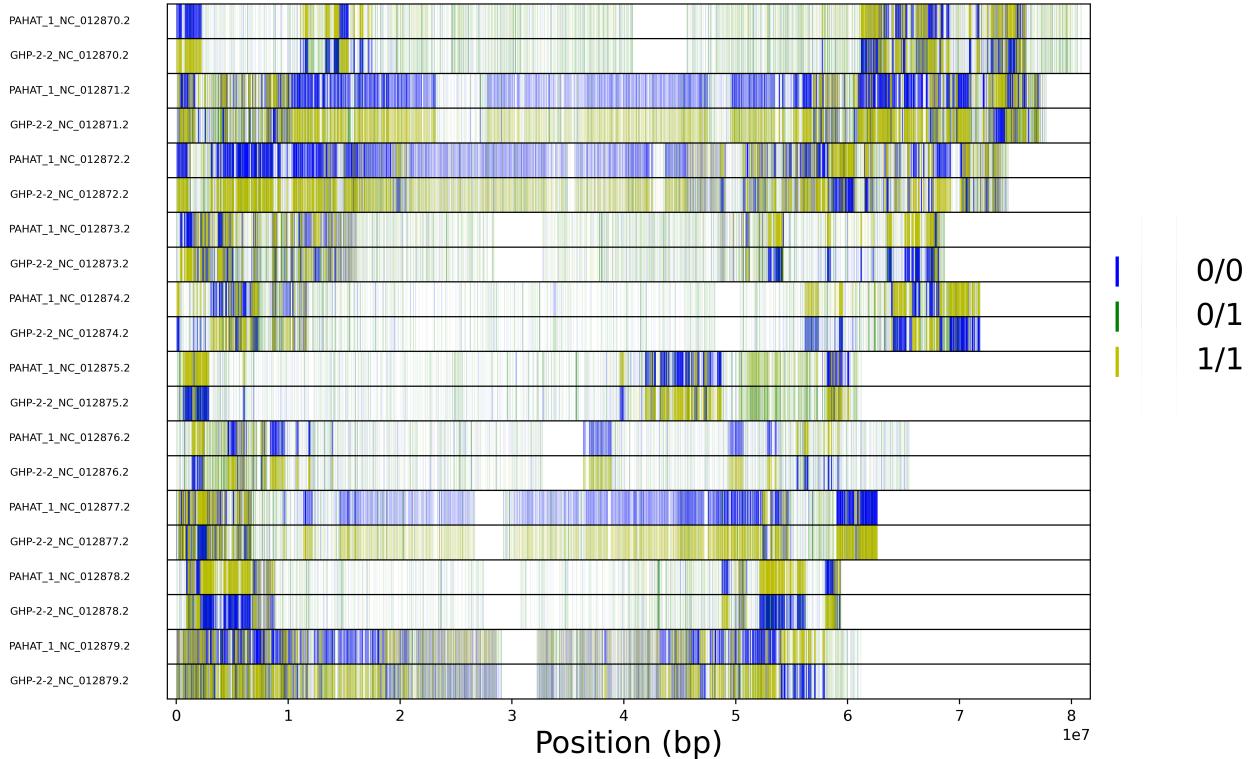
Contingency Table - Chromosome all

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	15850	69051	0
	0/1	8133	9738	4646	0
	1/1	48313	7313	0	0
	other	0	0	0	0

GT Plot after GT Filtering

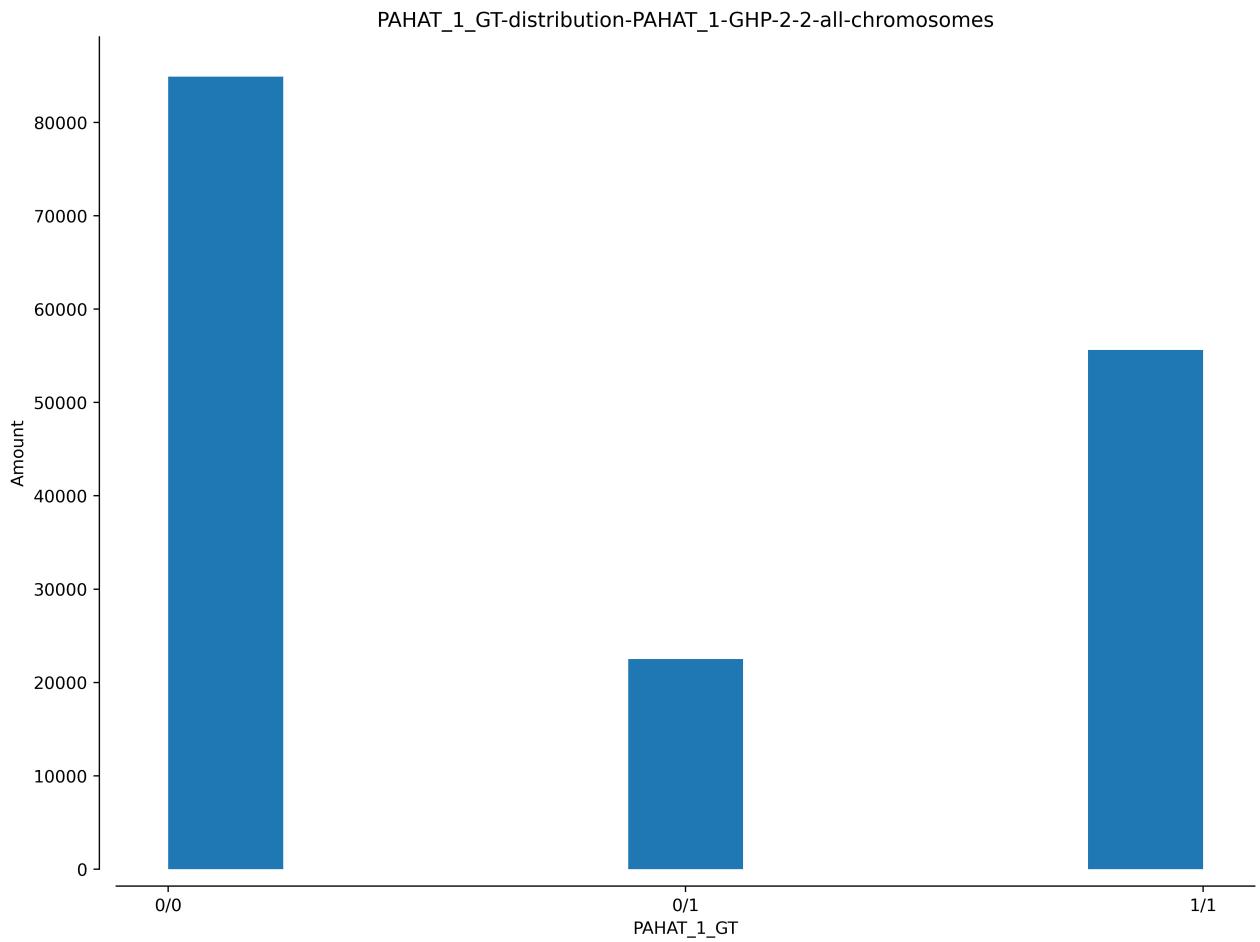
```
In [52]: plt.close('all')
GTplot(samples, vcf_df_04, chrom_len_00)
```

gt-plot-PAHAT_1-GHP-2-2

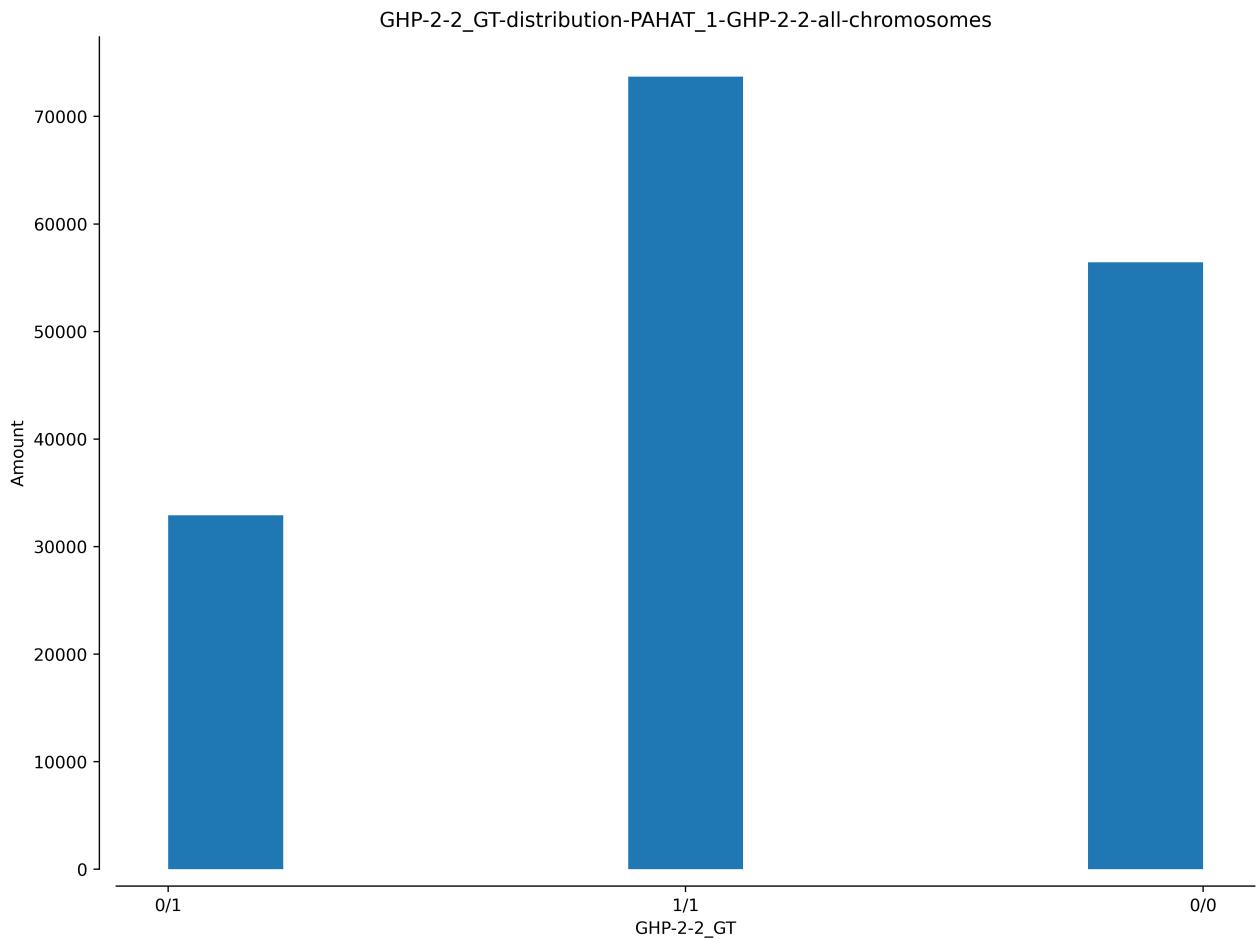


Histograms GT after GT Filtering

```
In [53]: plot_variant_hist(samples, vcf_df_04, 'all', 'PAHAT_1_GT', bins=9)
```



In [54]: `plot_variant_hist(samples, vcf_df_04, 'all', 'GHP-2-2_GT', bins=9)`



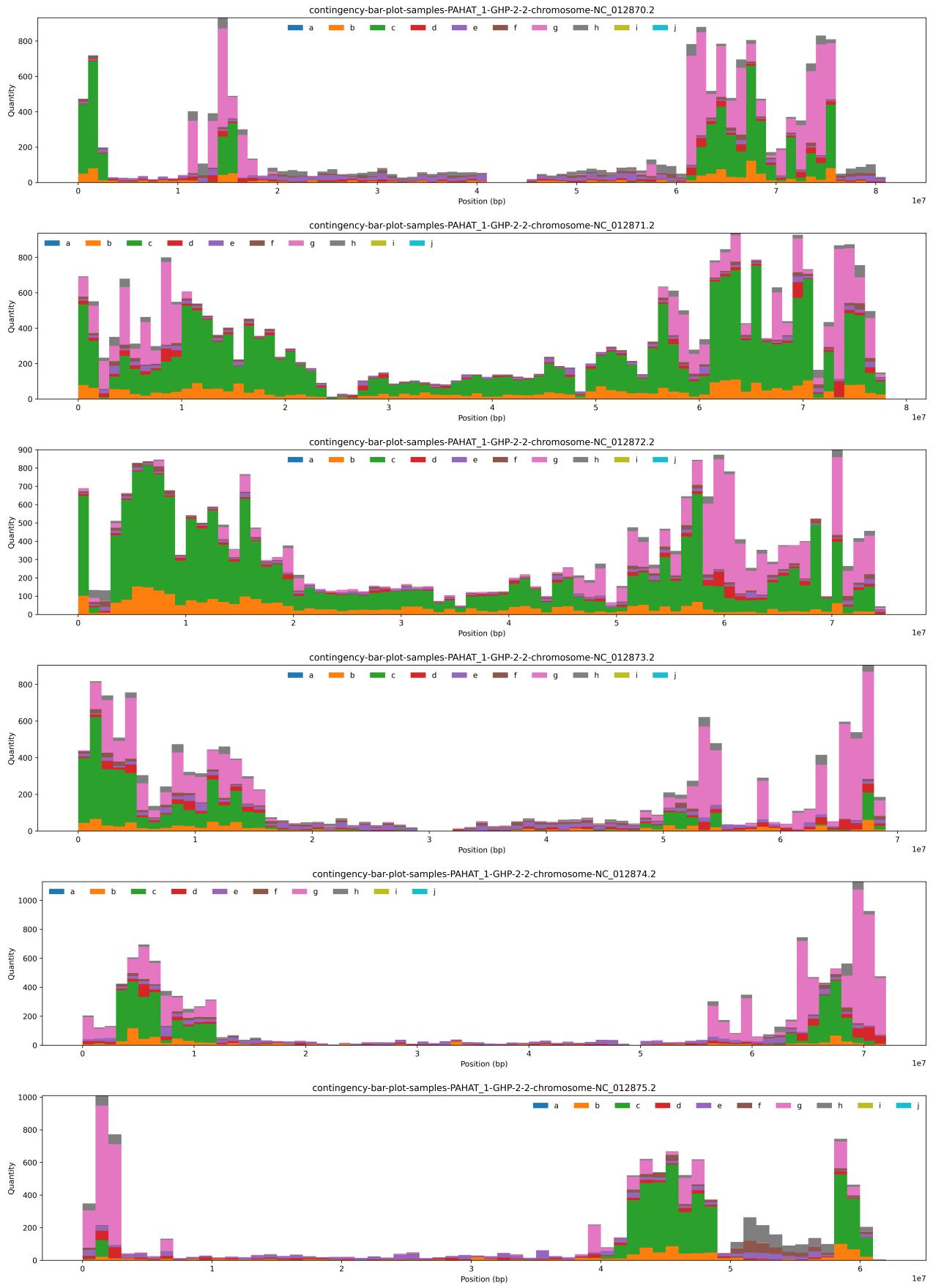
PART 6: Stacked Bar Plots

In [59]: `ct_guide()`

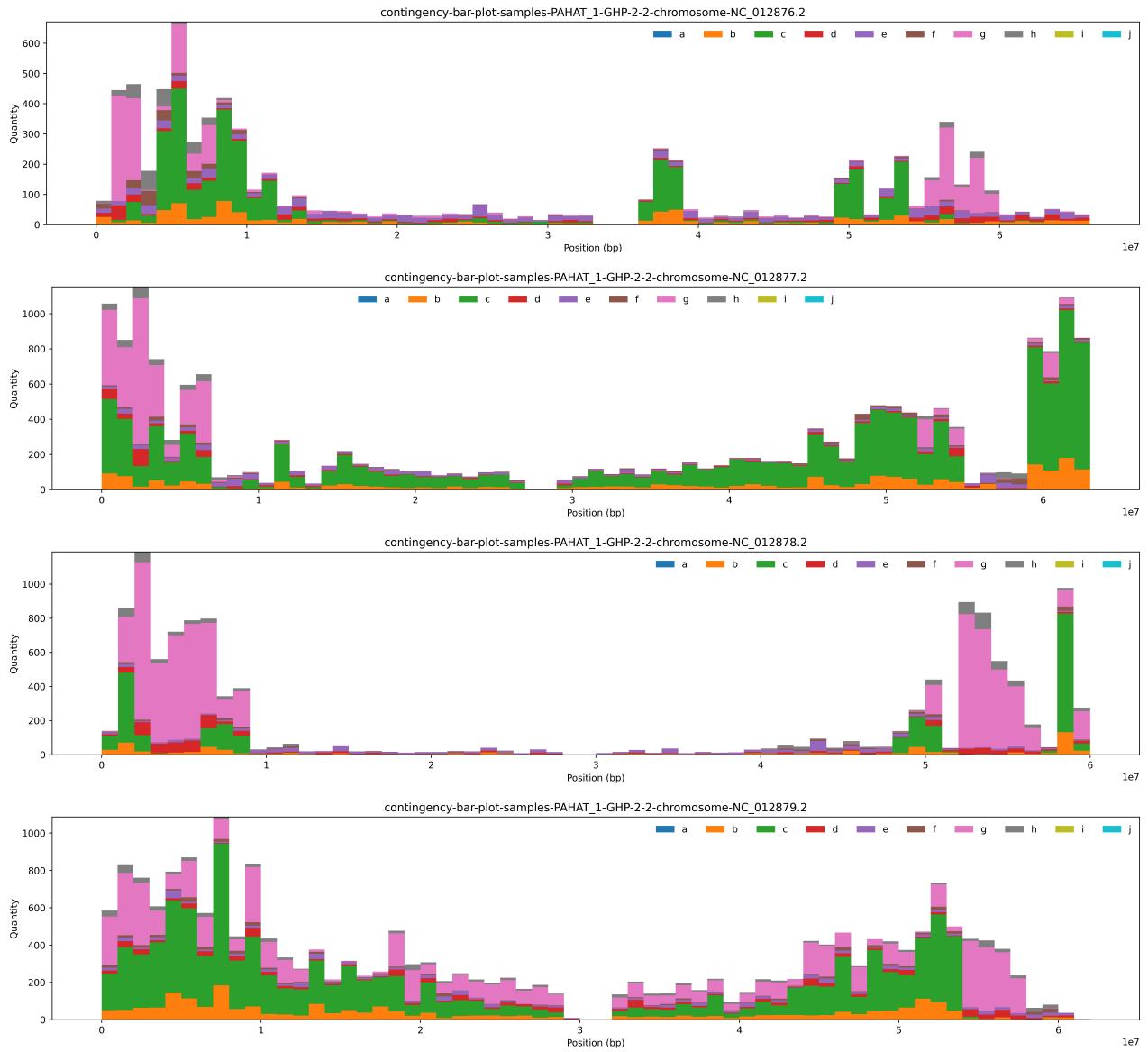
Out[59]:

		Mutant			
		0/0	0/1	1/1	other
Progenitor	0/0	a	b	c	
	0/1	d	e	f	
	1/1	g	h	i	
	other				j

In [55]: `plt.close('all')`
`window_size = 1000000`
`CTbarPlots(samples, vcf_df_04, chrom_len_00, window_size)`



Processing math: 100%

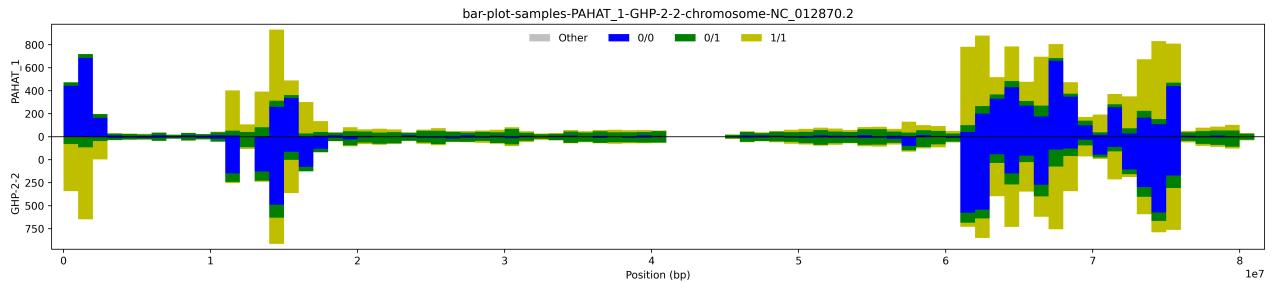


PART 7: Bar Plots per Chromosome

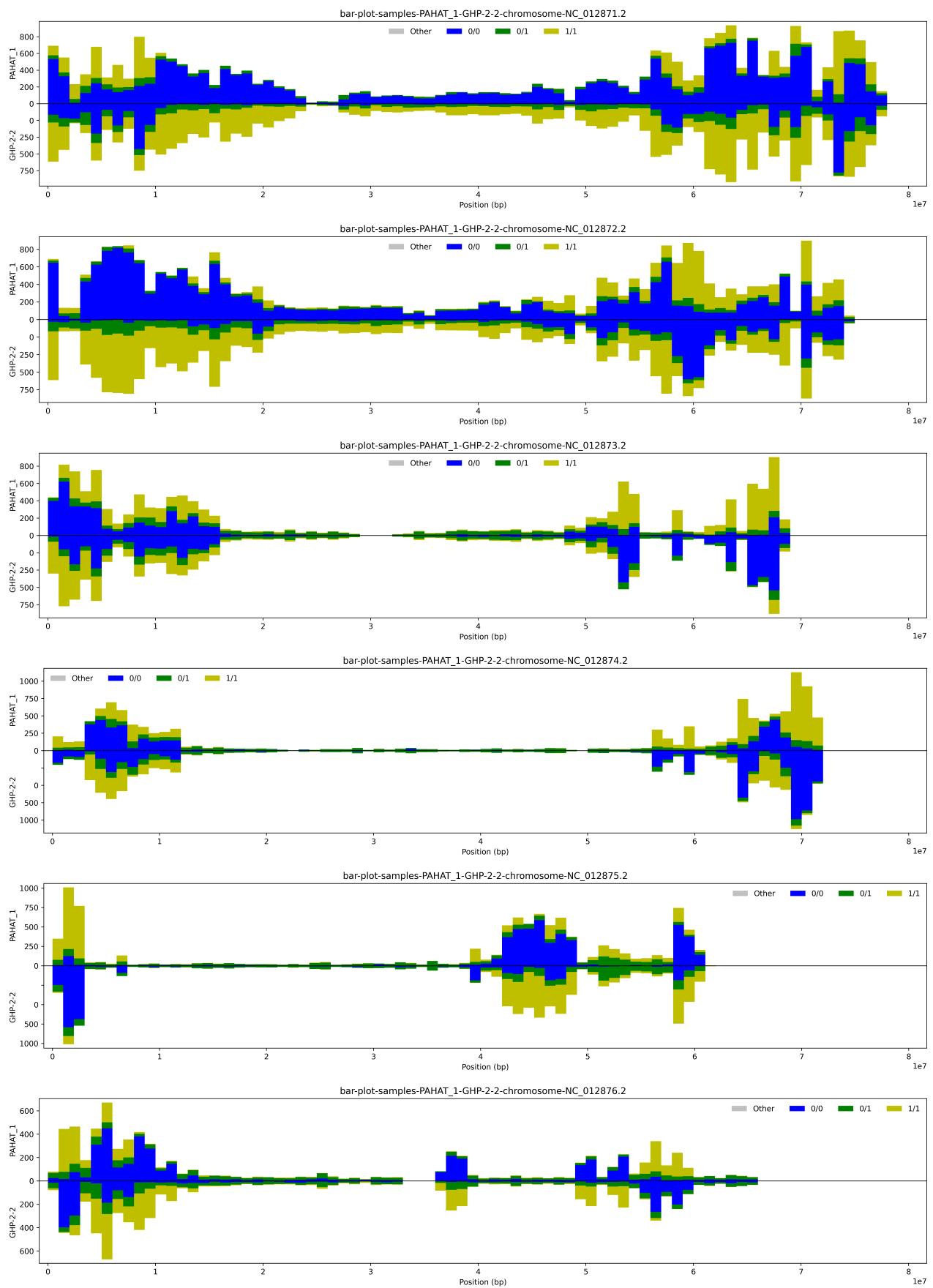
In [56]:

```
# suppress all the warnings from the inverted tickes of bar plots
import warnings
warnings.filterwarnings('ignore')

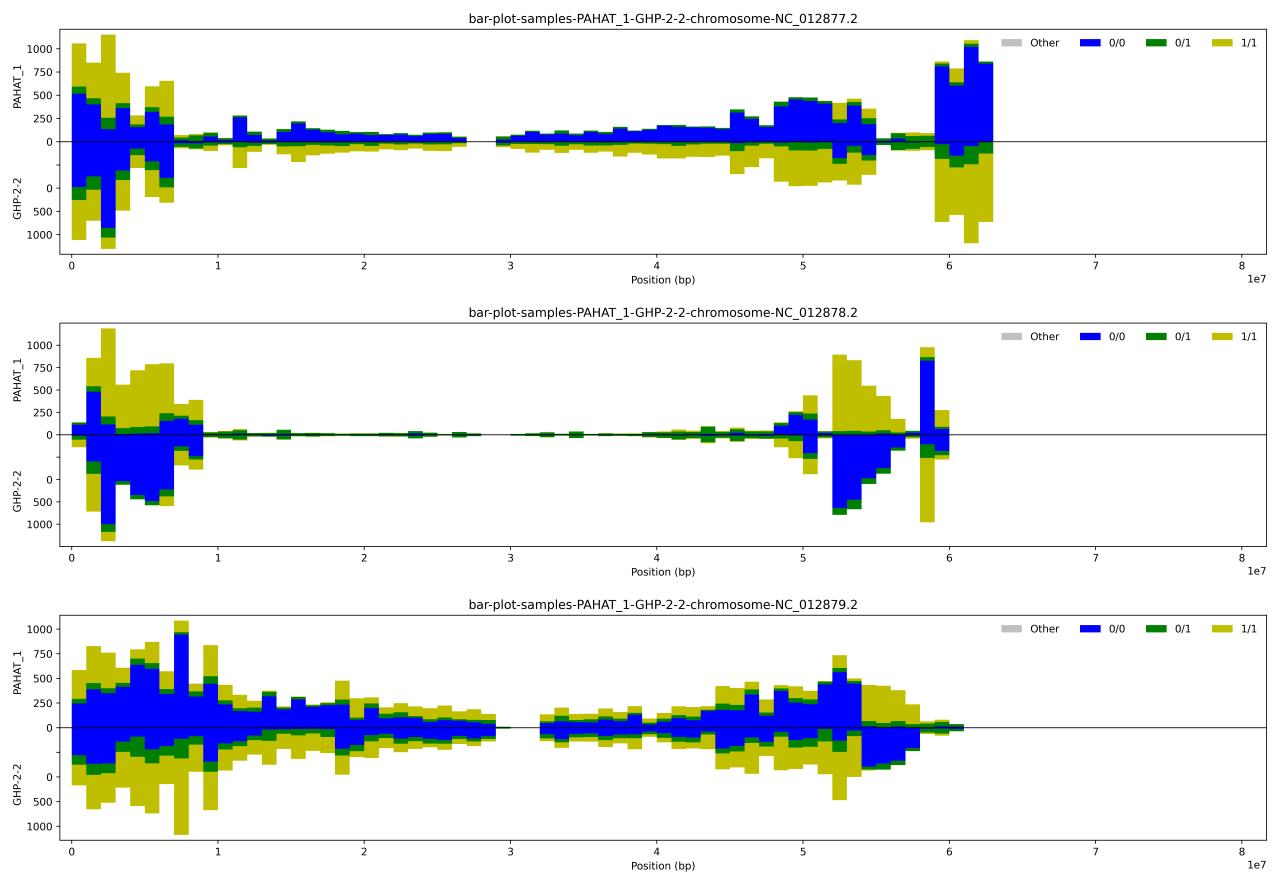
plt.close('all')
GTbarPlots(samples, vcf_df_04, chrom_len_00, window_size)
```



Processing math: 100%



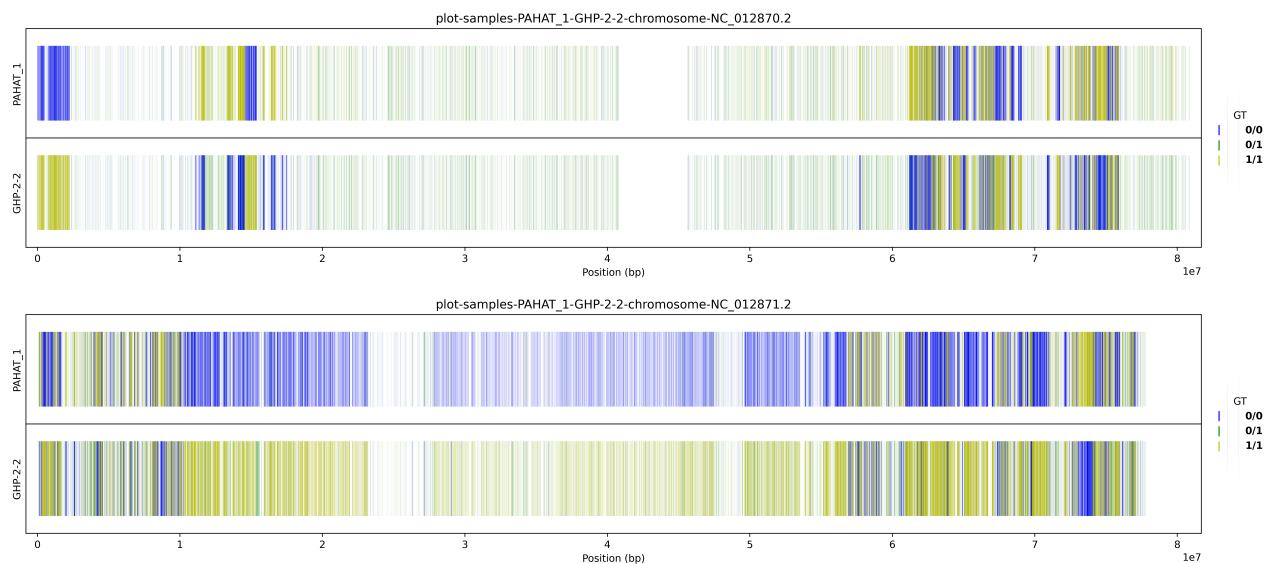
Processing math: 100%



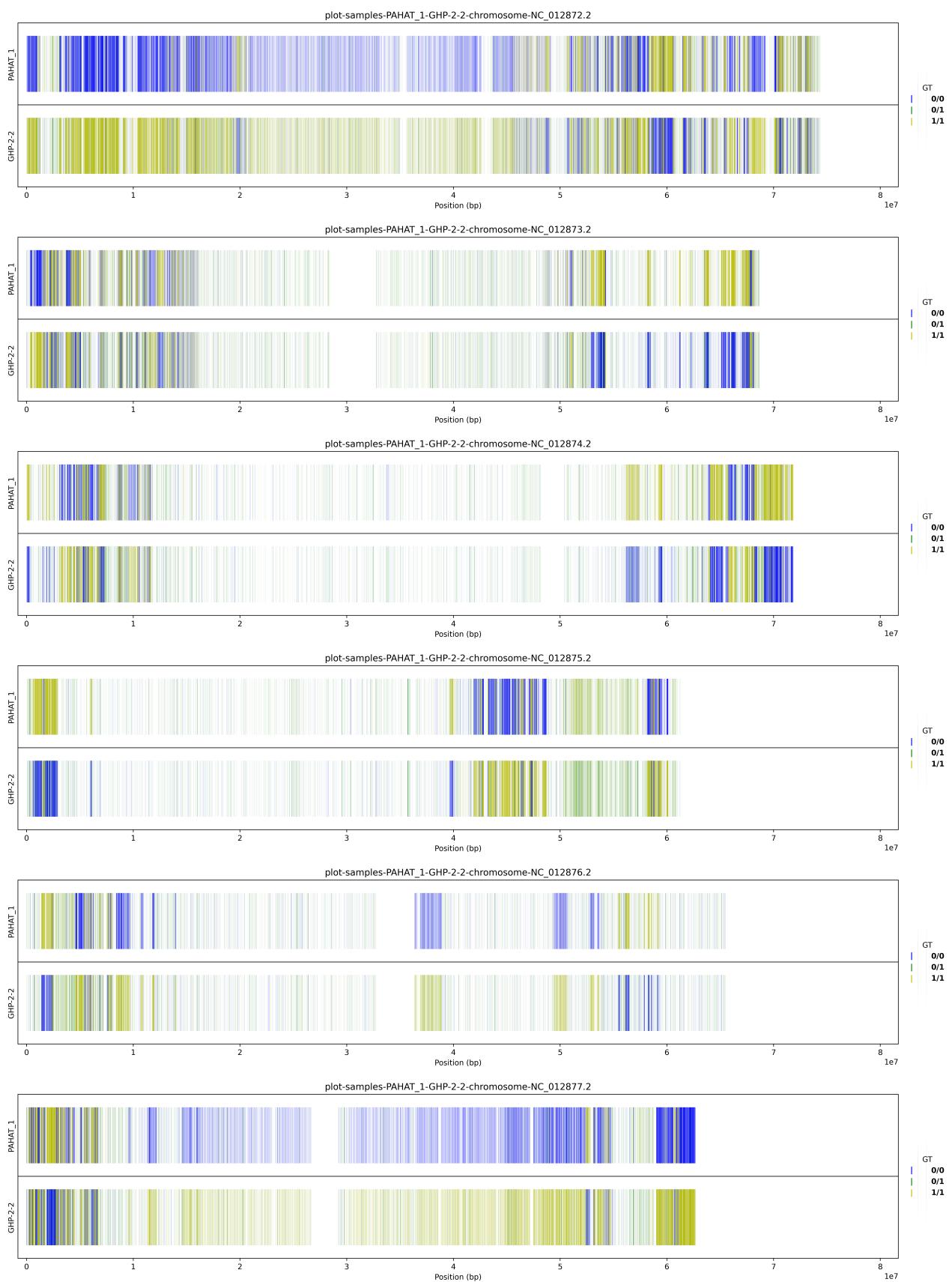
PART 8: GT Plots per Chromosome

In [57]:

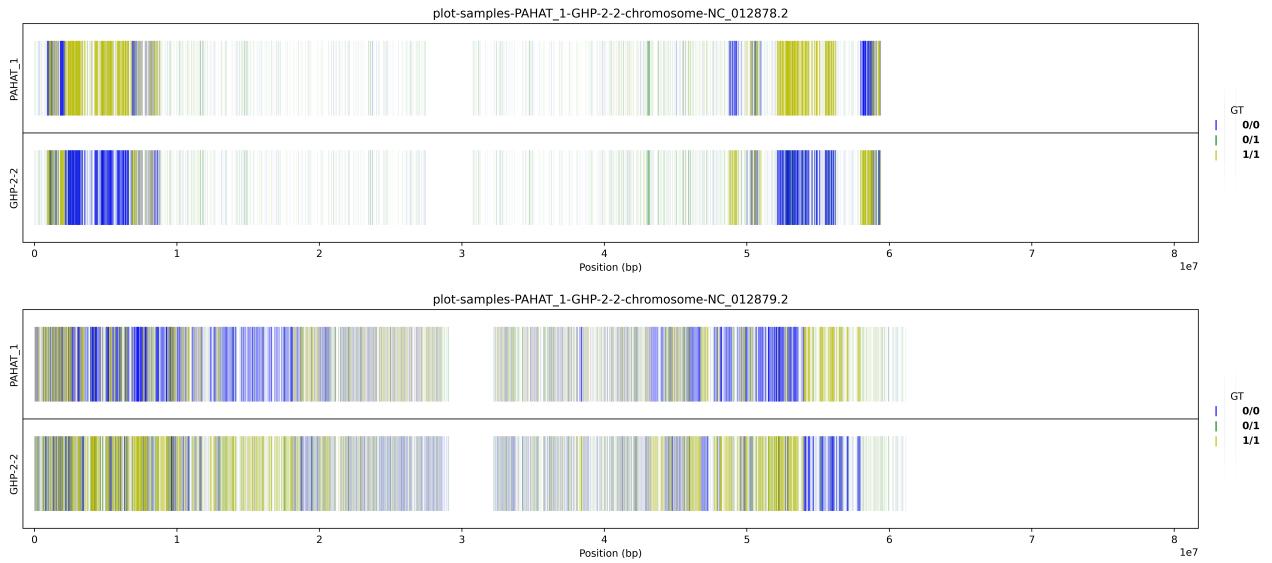
```
plt.close('all')
GTplots(samples, vcf_df_04, chrom_len_00)
```



Processing math: 100%



Processing math: 100%



PART 9: Contingency Table per Chromosome

In [58]:

```
import datafram_image as dfi

for chromosome in chrom_len_00.index:
    chromosome_df = vcf_df_04[ vcf_df_04.CHROM == chromosome ]

    # reset chromosome_df indexes for contingency table
    chromosome_df.reset_index(inplace=True, drop=True)
    chromosome_ct = contingency_table(samples, chromosome_df, chromosome)
```

Contingency Table - Chromosome NC_012870.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	1179	4670	0
	0/1	871	1278	635	0
	1/1	5900	1419	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012871.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	3090	15467	0
	0/1	1039	1010	656	0
	1/1	5268	733	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012872.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	3002	14417	0
	0/1	888	777	638	0
	1/1	5683	685	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012873.2

Processing math: 100%

		GHP-2-2_GT			
		0/0	0/1	1/1	other

PAHAT_1_GT	0/0	0	946	3526	0
	0/1	901	1090	551	0
	1/1	5780	882	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012874.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	850	2945	0
	0/1	894	1206	302	0
	1/1	5307	564	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012875.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	905	3644	0
	0/1	599	1049	509	0
	1/1	2659	855	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012876.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	873	2815	0
	0/1	553	988	261	0
	1/1	1819	365	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012877.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	2006	9975	0
	0/1	603	856	449	0
	1/1	3038	414	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012878.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	708	2152	0
	0/1	811	711	208	0
	1/1	6689	714	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_012879.2

		GHP-2-2_GT			
		0/0	0/1	1/1	other
PAHAT_1_GT	0/0	0	2291	9440	0
	0/1	974	773	437	0
	1/1	6170	682	0	0
	other	0	0	0	0