

Cowpea - SNPs 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_cowpea = '/home/anibal/genome_files/freebayes~bwa~GCF_004118075.1_ASM411807v1.200.vcf.gz'
```

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

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

```
Out[4]: array(['CBC1_P1', 'CBC5_A1'], dtype=object)
```

```
In [5]: progenitor = 'CBC1_P1'
mutant = 'CBC5_A1'
samples = [progenitor, mutant]
samples
```

```
Out[5]: ['CBC1_P1', 'CBC5_A1']
```

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

```
Out[6]:
```

	CHROM	POS	REF
0	NC_018051.1	11786	T
1	NC_018051.1	11801	TCTTCCT
2	NC_018051.1	11813	AGCC
3	NC_018051.1	11825	GGTAGGTAAT
4	NC_018051.1	18327	A
...
1968687	NC_040289.1	41659114	T
1968688	NC_040289.1	41659137	G
1968689	NC_040289.1	41667130	GTTTCA

	CHROM	POS	REF
1968690	NC_040289.1	41667148	T
1968691	NC_040289.1	41668013	CAGGGTTAGGGTTAGGGTTCAGGGTTAGGGTTAGGGTTCAGG...

1968692 rows × 14 columns

◀ ▶

In [7]: chrom_len

Out[7]: LEN

CHROM	LEN
NC_040279.1	42129361
NC_040280.1	33908088
NC_040281.1	65292630
NC_040282.1	42731077
NC_040283.1	48746289
NC_040284.1	34463471
NC_040285.1	40876636
NC_040286.1	38363498
NC_040287.1	43933251
NC_040288.1	41327797
NC_040289.1	41684185
NC_018051.1	152415

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')

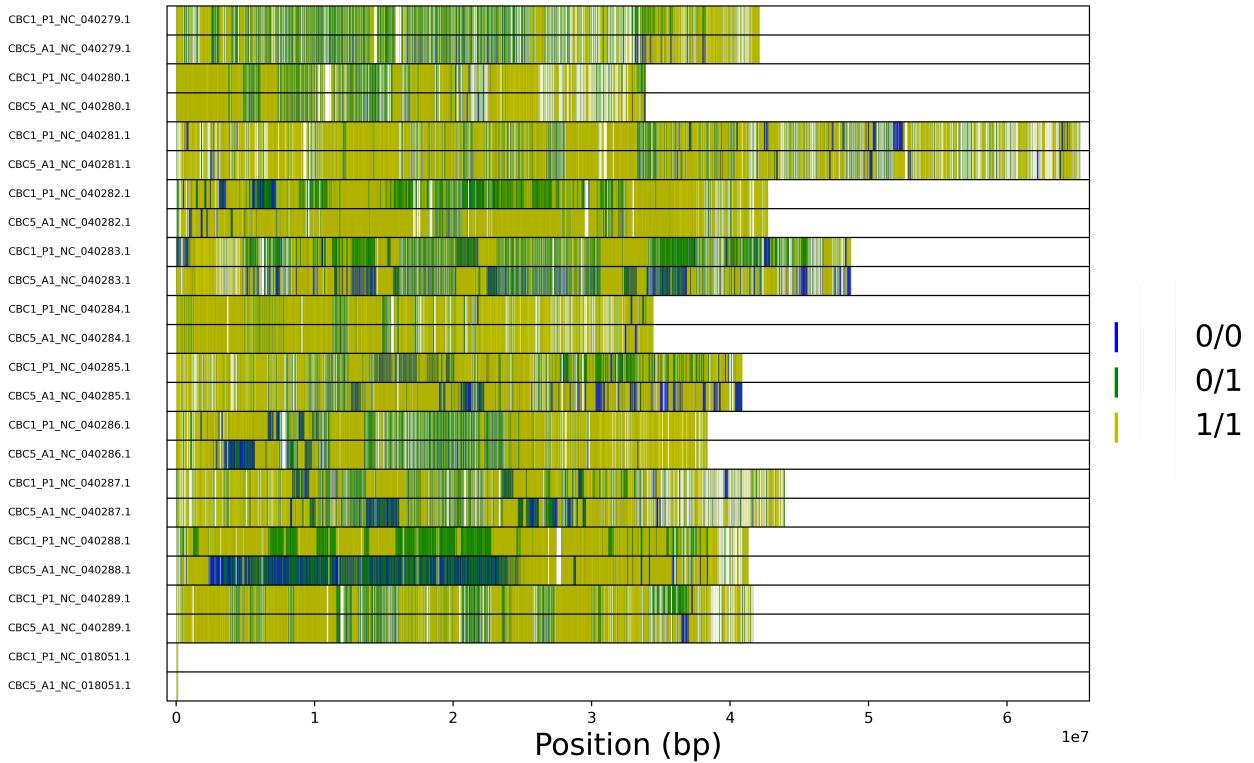
Contingency Table - Chromosome all

CBC1_P1_GT	CBC5_A1_GT			
	0/0	0/1	1/1	other
0/0	0	52496	132507	45048
0/1	287090	273476	178974	45048
1/1	211458	19728	767915	45048
other	45048	45048	45048	45048

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

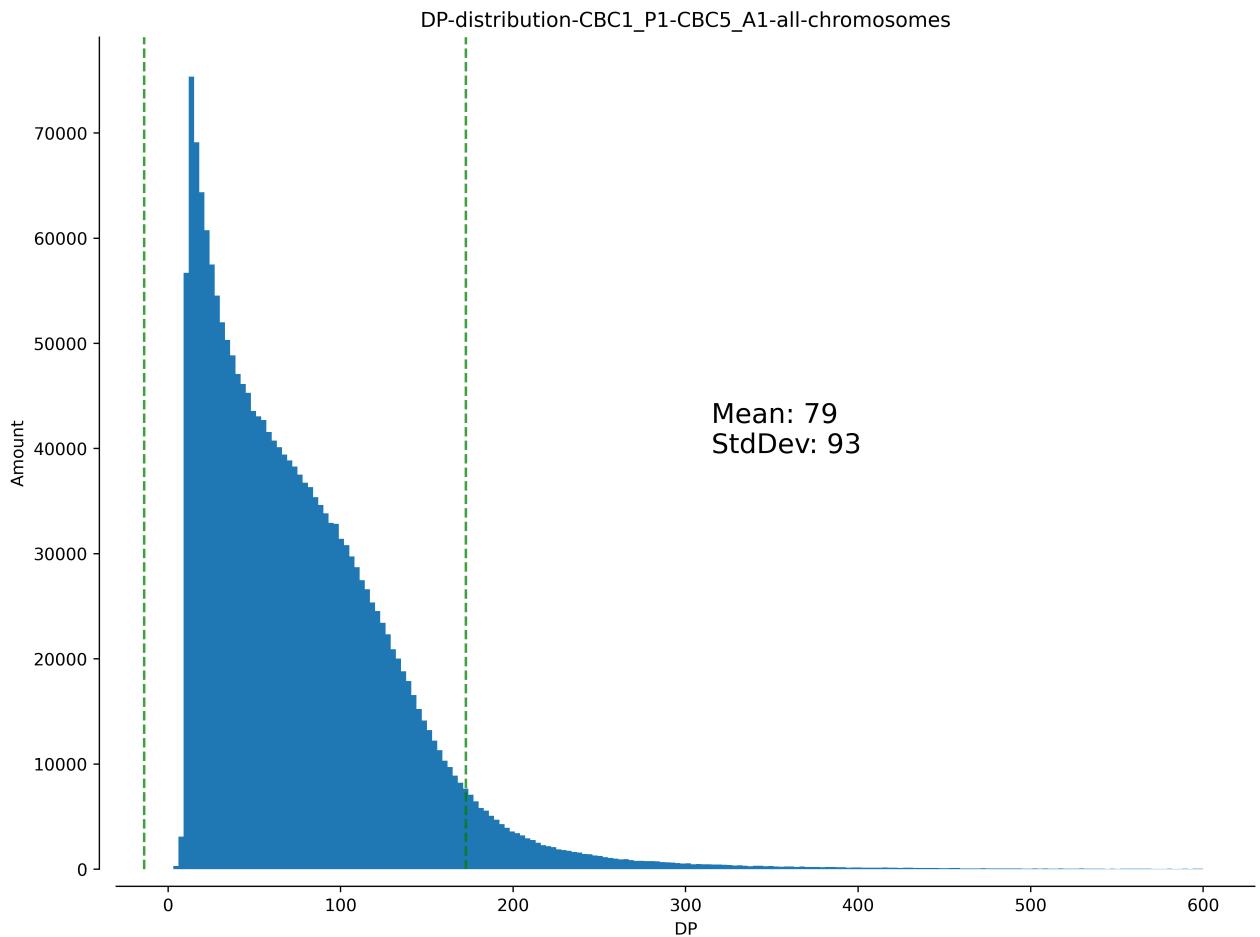
In [9]: plt.close('all')
Gtplot(samples, vcf_df, chrom_len)

gt-plot-CBC1_P1-CBC5_A1

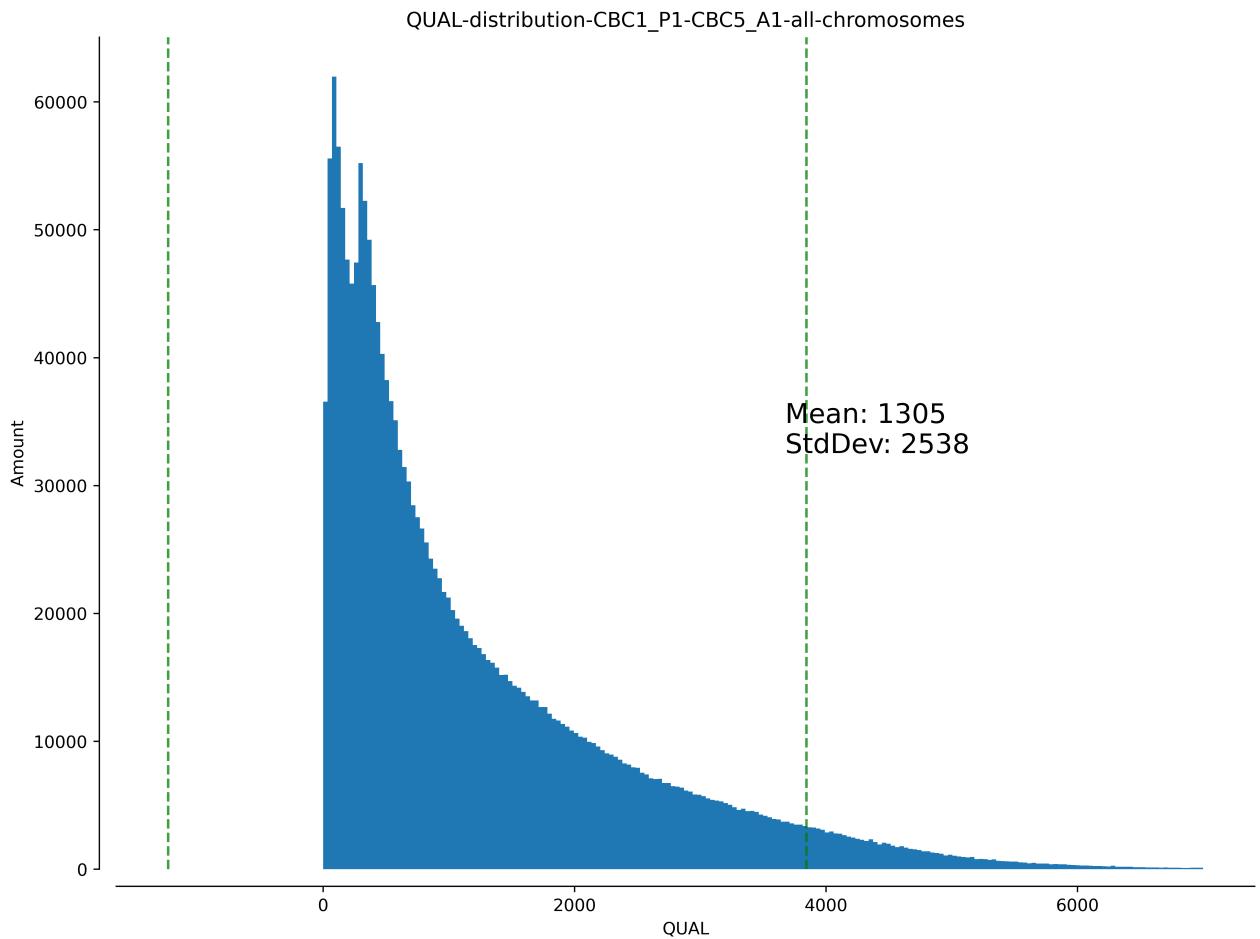


Histograms - DP , QUAL , TYPE and GT Attributes - All Chromosomes - Unfiltered

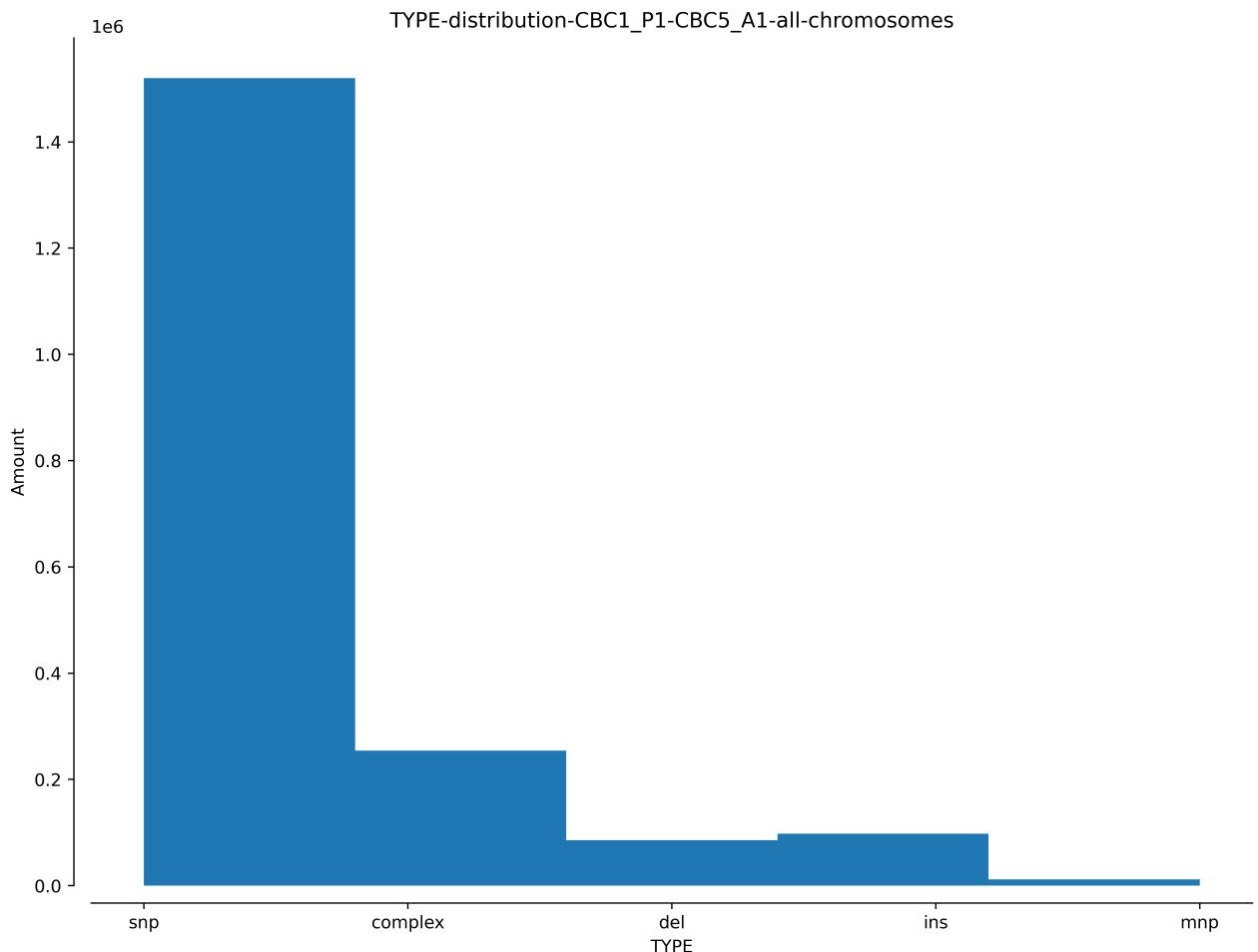
```
In [10]: plot_variant_hist(samples, vcf_df, 'all', 'DP', bins=200, MSTD=True, xmax=600)
```



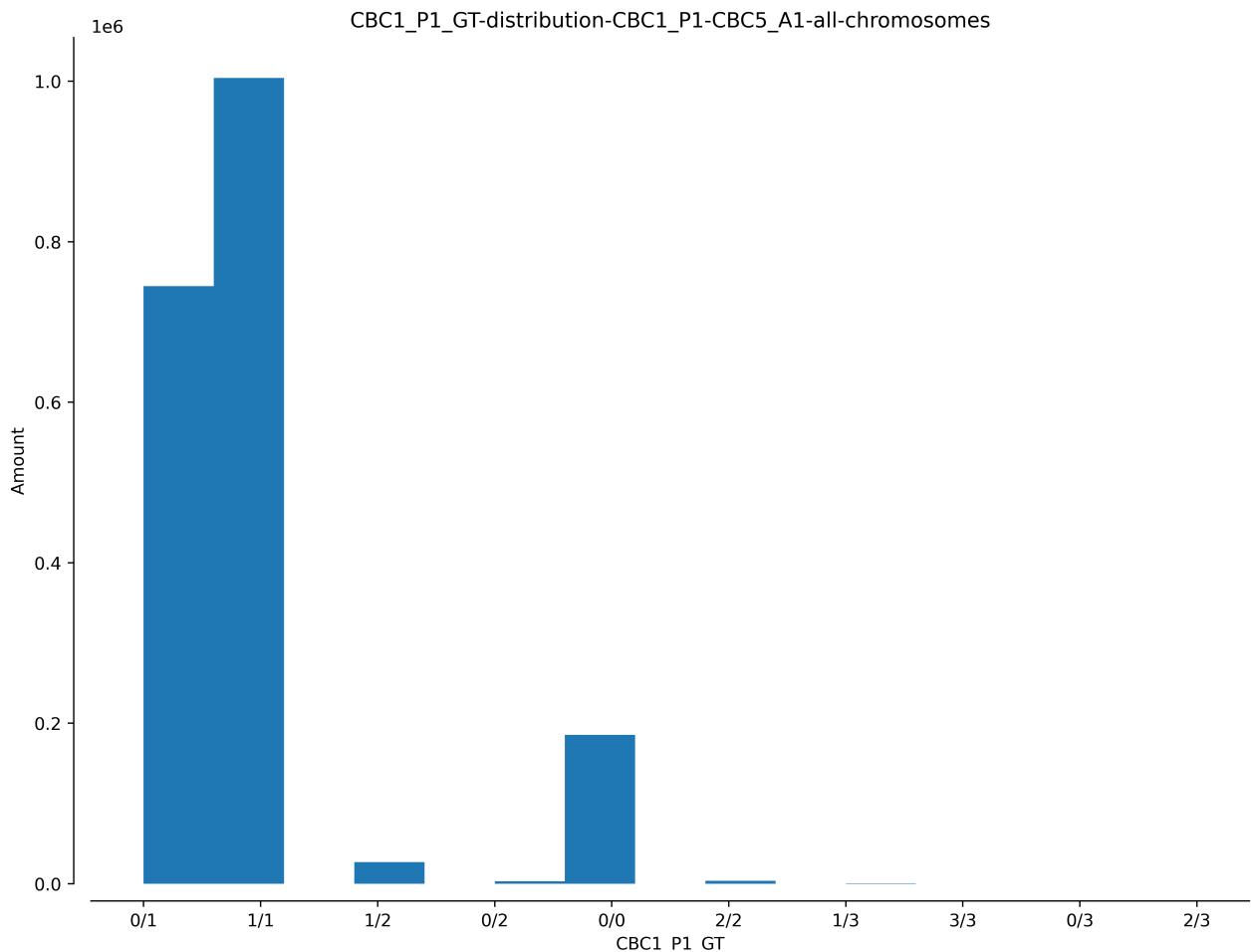
```
In [11]: plot_variant_hist(samples, vcf_df, 'all', 'QUAL', bins=200, MSTD=True, xmax=700)
```



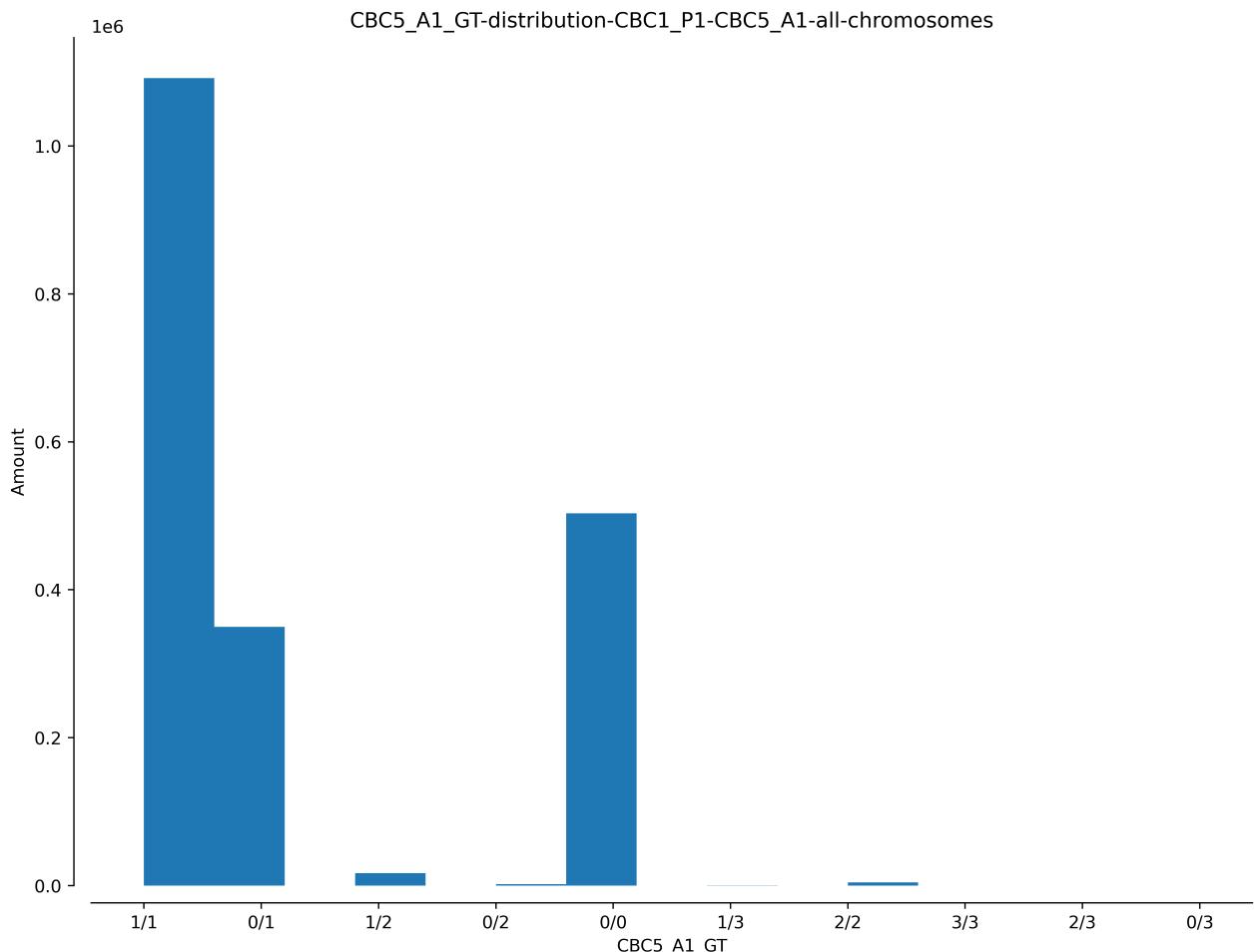
In [12]: `plot_variant_hist(samples, vcf_df, 'all', 'TYPE', bins=5)`



```
In [14]: plot_variant_hist(samples, vcf_df, 'all', 'CBC1_P1_GT', bins=15)
```



```
In [15]: plot_variant_hist(samples, vcf_df, 'all', 'CBC5_A1_GT', bins=15)
```



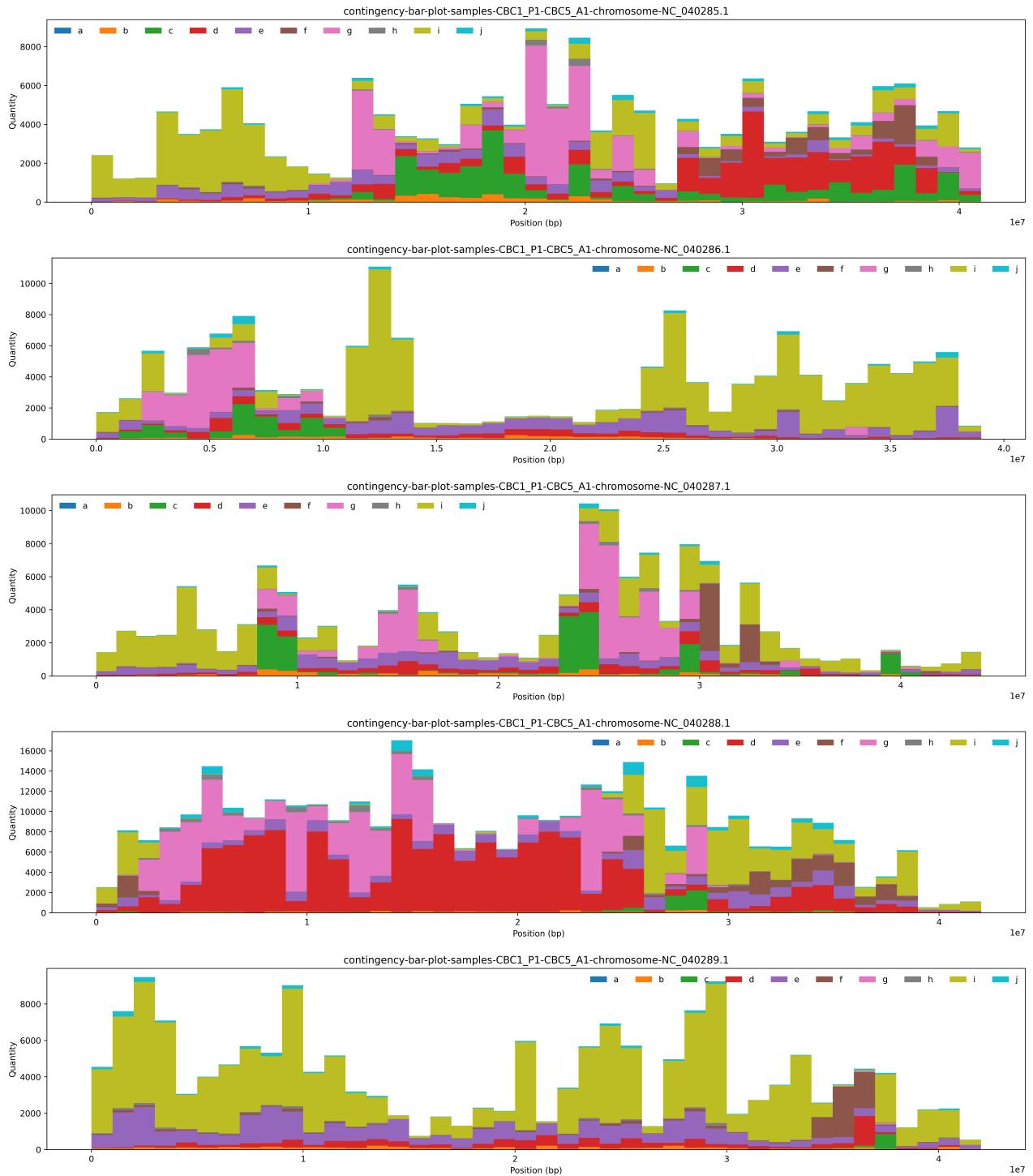
Stacked Bar Plots - RAW

```
In [16]: ct_guide()
```

		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 [17]: 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 [18]:

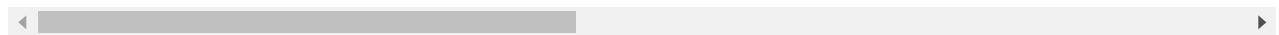
```
drop_mito_chloro = "CHROM != NC_018051.1"
vcf_df_00 = filter_vcf(vcf_df, drop_mito_chloro)
vcf_df_00
```

Out[18]:

CHROM	POS	REF
-------	-----	-----

	CHROM	POS	REF
0	NC_040279.1	912	G
1	NC_040279.1	948	AGGGGAAAC A
2	NC_040279.1	1173	C
3	NC_040279.1	1390	C
4	NC_040279.1	1424	T
...
1968602	NC_040289.1	41659114	T
1968603	NC_040289.1	41659137	G
1968604	NC_040289.1	41667130	GTTTCA
1968605	NC_040289.1	41667148	T
1968606	NC_040289.1	41668013	CAGGGTTAGGGTTAGGGTTCAGGGTTAGGGTTAGGGTTCAGG...

1968607 rows × 14 columns



```
In [19]: mito_chloro = ['NC_018051.1']
chrom_len_00 = chrom_len.drop(mito_chloro)
chrom_len_00
```

Out[19]: LEN

CHROM	LEN
NC_040279.1	42129361
NC_040280.1	33908088
NC_040281.1	65292630
NC_040282.1	42731077
NC_040283.1	48746289
NC_040284.1	34463471
NC_040285.1	40876636
NC_040286.1	38363498
NC_040287.1	43933251
NC_040288.1	41327797
NC_040289.1	41684185

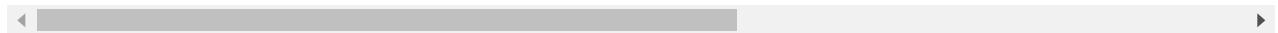
Create Mitochondria and Chloroplast Variants and Chomosome Length Dataframes

```
In [20]: drop_chrom = "CHROM!=NC_040279.1, CHROM!=NC_040280.1, CHROM!=NC_040281.1, CHROM!=NC_040282.1, CHROM!=NC_040283.1, CHROM!=NC_040284.1, CHROM!=NC_040285.1, CHROM!=NC_040286.1, CHROM!=NC_040287.1, CHROM!=NC_040288.1, CHROM!=NC_040289.1"
vcf_df_mito_chloro = filter_vcf(vcf_df, drop_chrom)
vcf_df_mito_chloro
```

Out[20]:

	CHROM	POS	REF	ALT	QUAL	DP	CBC1_P1_DP	CBC5_A1_DF
0	NC_018051.1	11786	T	C	492.441010	22	9	13
1	NC_018051.1	11801	TCTTCCT	CCTACCC	319.031006	26	10	16
2	NC_018051.1	11813	AGCC	GGCT	305.860992	28	10	18
3	NC_018051.1	11825	GGTAGGTAAT	AGTGGGGAAC	315.924988	28	9	19
4	NC_018051.1	18327	G	A	580.976990	25	15	10
...
80	NC_018051.1	129428	T	C	245.087997	20	10	10
81	NC_018051.1	132082	C	G	586.130981	33	16	17
82	NC_018051.1	132156	C	T	45.394600	19	10	9
83	NC_018051.1	132169	CCGGT	ACGGG	478.346985	19	9	10
84	NC_018051.1	132184	A	G	21.932899	19	9	10

85 rows × 14 columns



In [21]:

```
mito_chloro_len = chrom_len.loc[mito_chloro]
mito_chloro_len
```

Out[21]:

LEN	CHROM
NC_018051.1	152415

Contingency Table - No Mitochondria/Chloroplast

In [22]:

```
contingency_table_1 = contingency_table(samples, vcf_df_00, 'all')
```

Contingency Table - Chromosome all

CBC1_P1_GT	CBC5_A1_GT			
	0/0	0/1	1/1	other
0/0	0	52496	132507	45041
0/1	287090	273450	178971	45041
1/1	211458	19719	767875	45041
other	45041	45041	45041	45041

GT Plot - No Mitochondria/Chloroplast

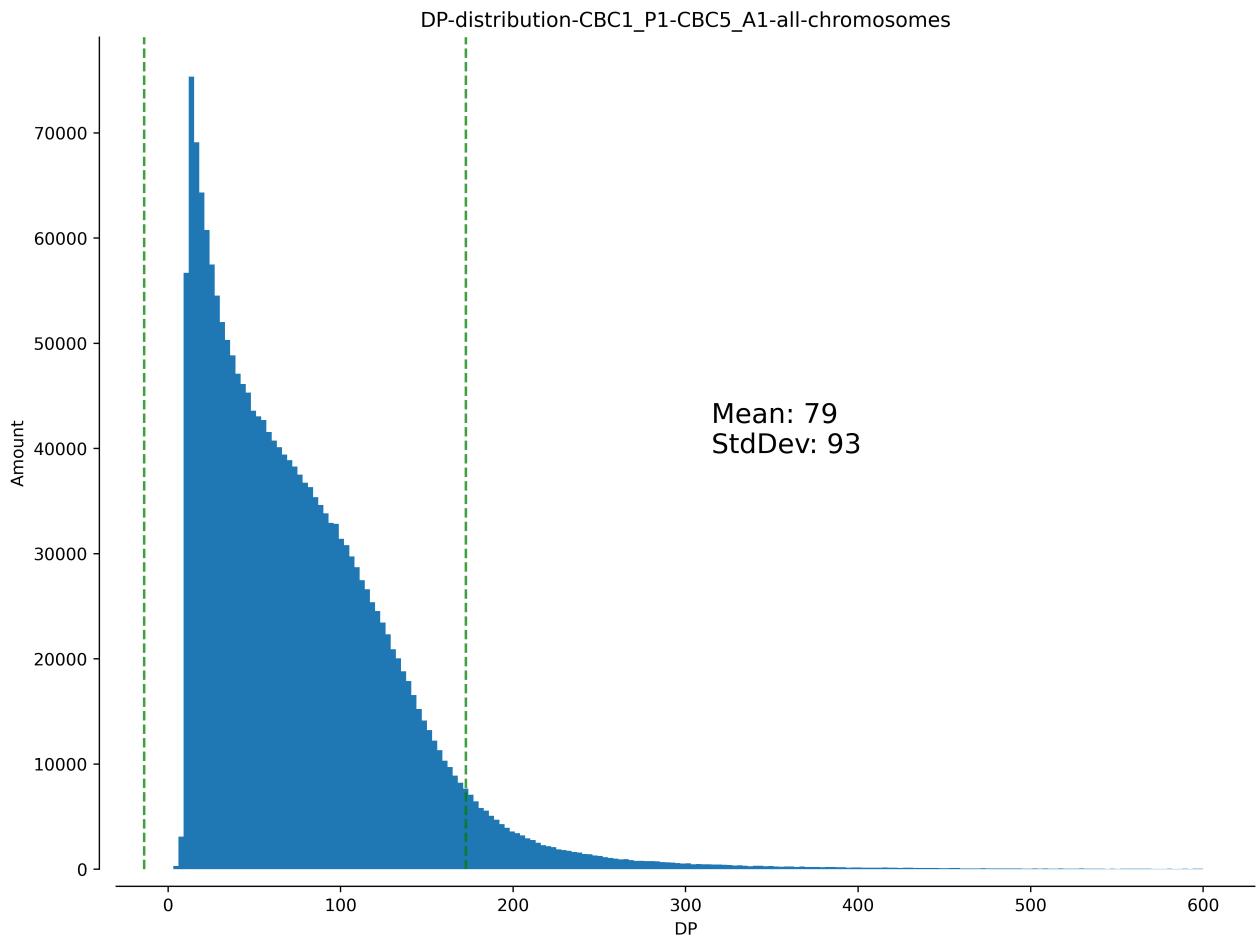
In [23]:

```
# plt.close('all')
# GTplot(samples, vcf_df_00, chrom_len_00)
```

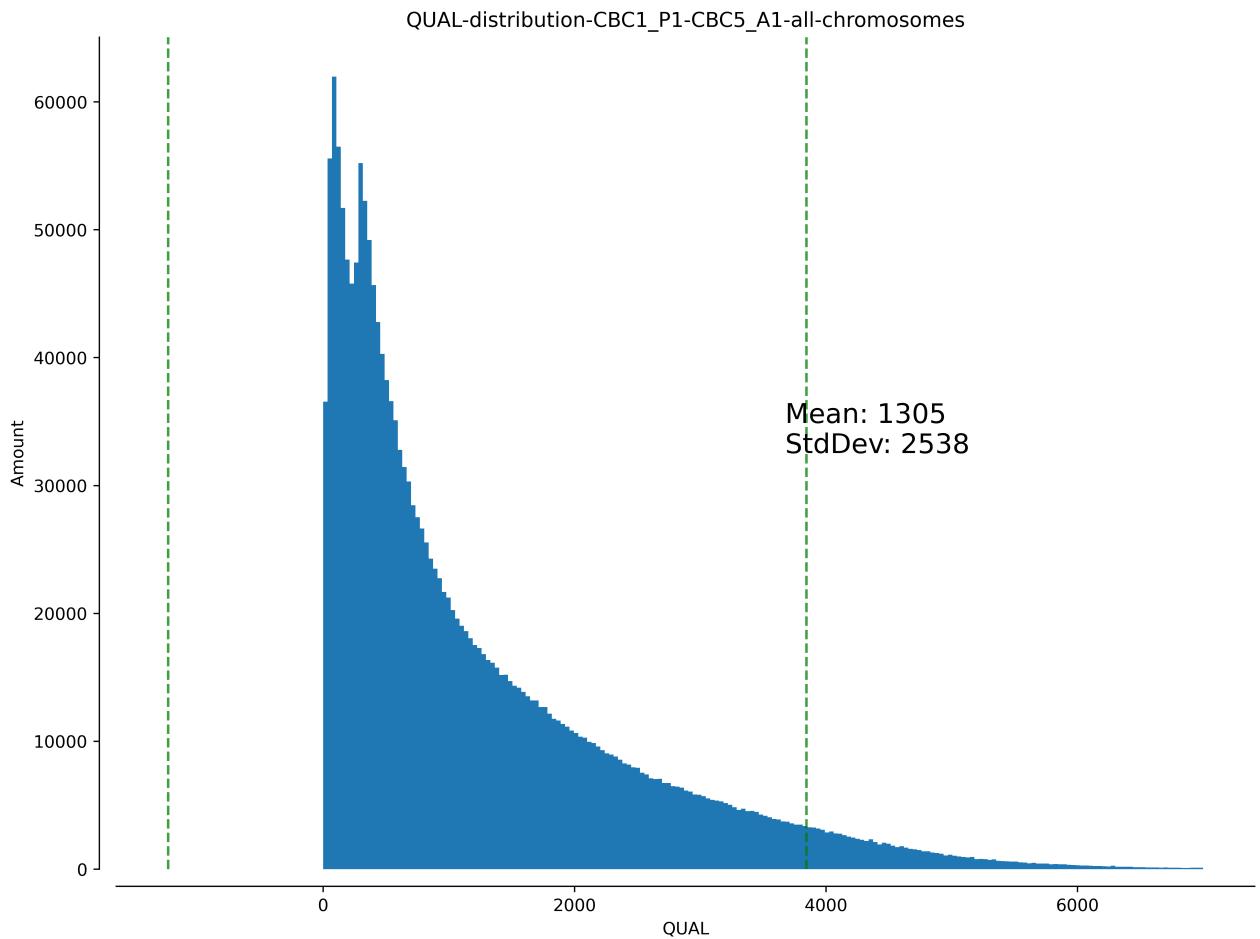
Histograms - DP , QUAL , TYPE and GT Attributes - No Mitochondria/Chloroplast

In [24]:

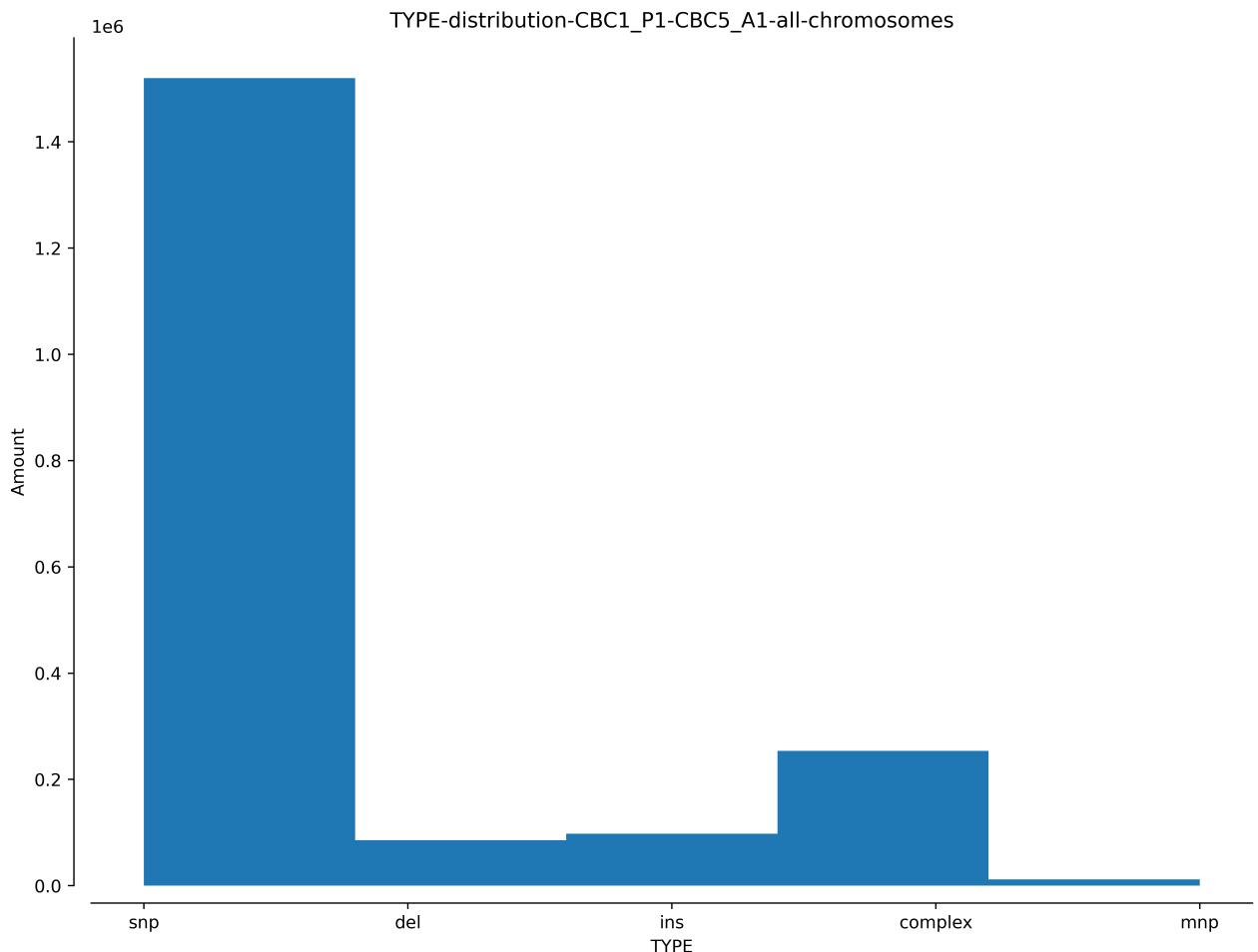
```
plot_variant_hist(samples, vcf_df_00, 'all', 'DP', bins=200, MSTD=True, xmax=600)
```



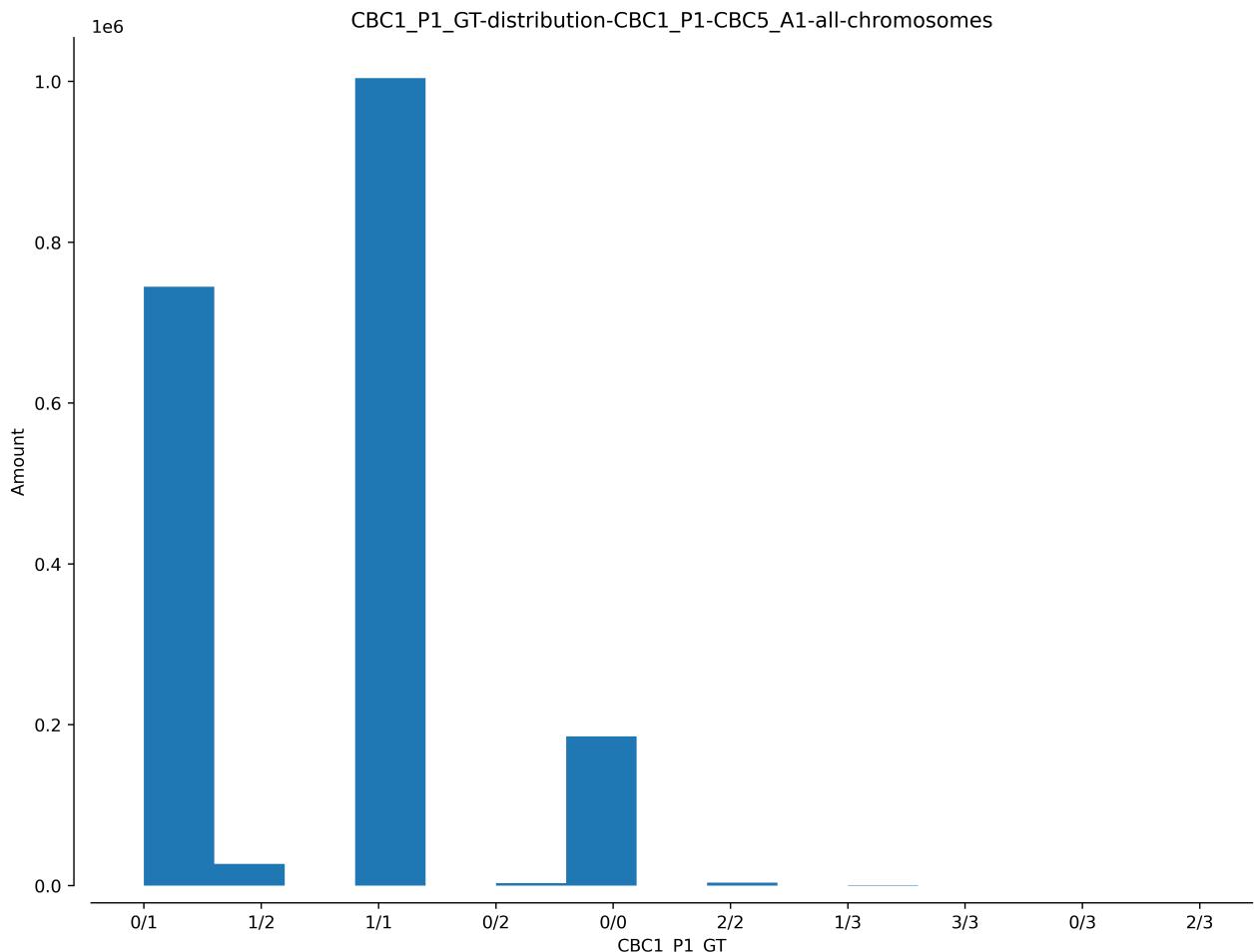
```
In [25]: plot_variant_hist(samples, vcf_df_00, 'all', 'QUAL', bins=200, MSTD=True, xmax=
```



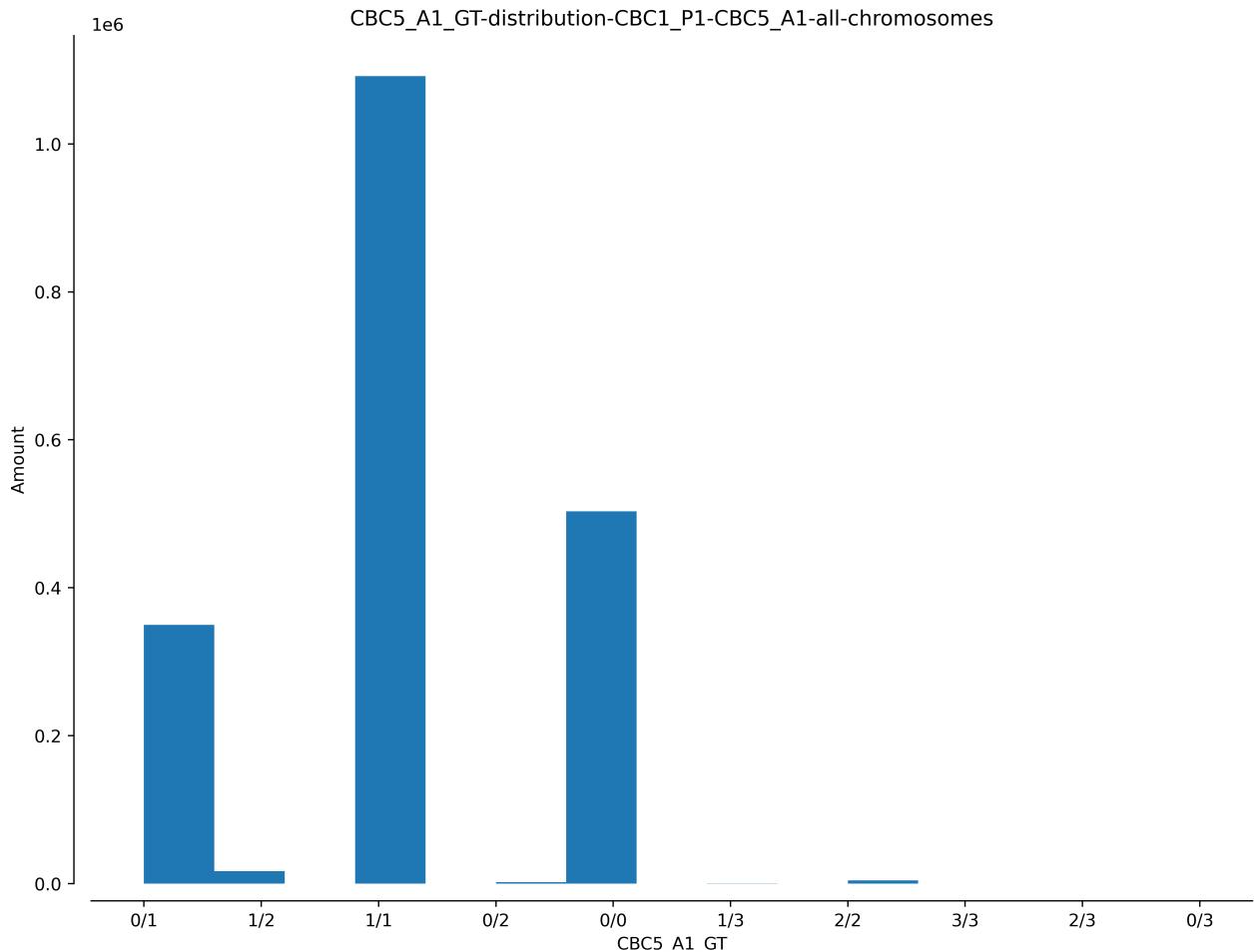
In [26]: `plot_variant_hist(samples, vcf_df_00, 'all', 'TYPE', bins=5)`



```
In [27]: plot_variant_hist(samples, vcf_df_00, 'all', '%s_GT' % progenitor, bins=15)
```



```
In [28]: plot_variant_hist(samples, vcf_df_00, 'all', '%s_GT' % mutant, bins=15)
```



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

In [29]:

```
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
```

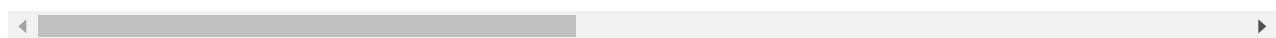
DP >= -107, DP <= 265

Out[29]:

	CHROM	POS	REF
0	NC_040279.1	912	G
1	NC_040279.1	948	AGGGGAAAC A
2	NC_040279.1	1173	C
3	NC_040279.1	1390	C
4	NC_040279.1	1424	T
...
1942510	NC_040289.1	41659114	T

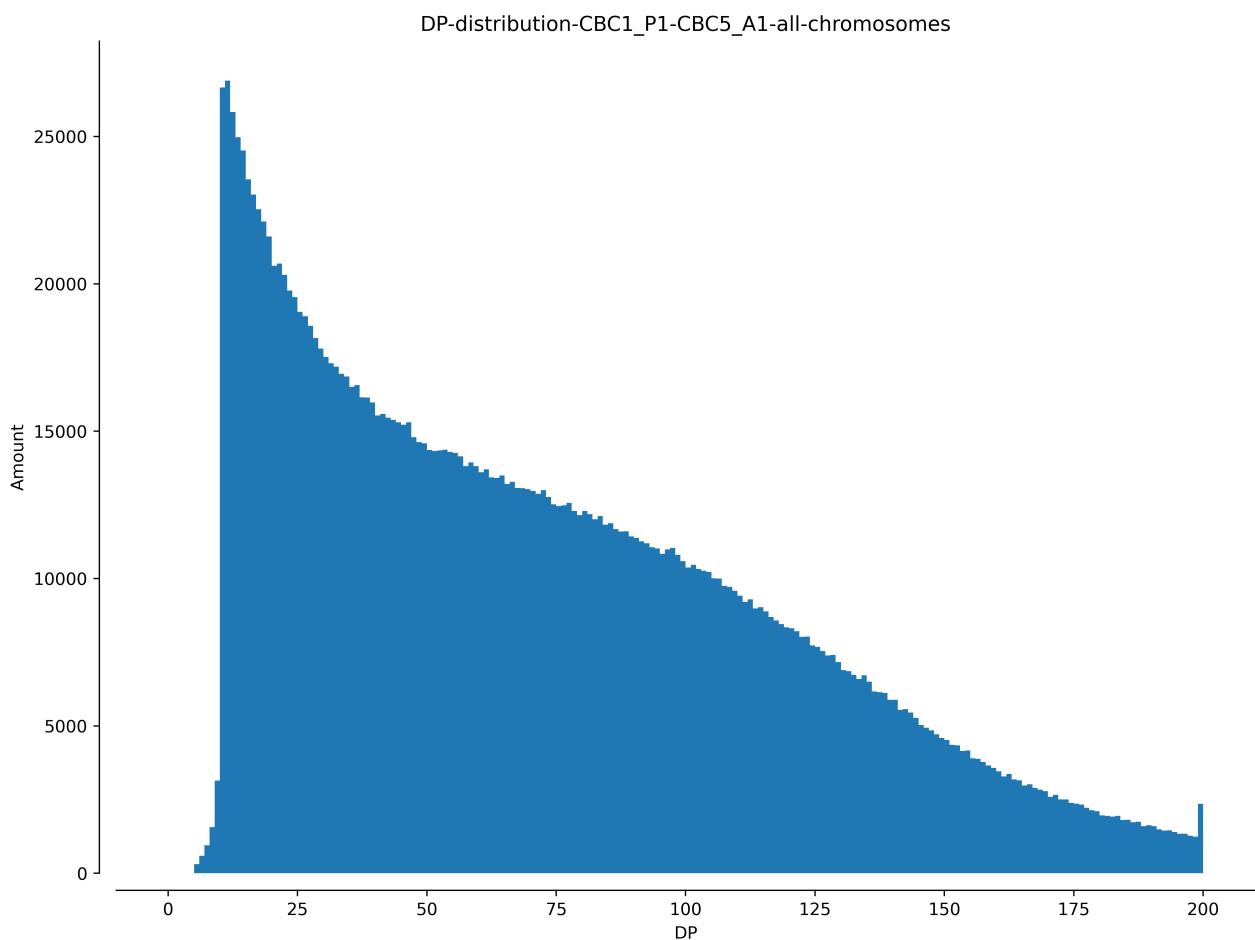
	CHROM	POS	REF
1942511	NC_040289.1	41659137	G
1942512	NC_040289.1	41667130	GTTTCA
1942513	NC_040289.1	41667148	T
1942514	NC_040289.1	41668013	CAGGGTTAGGGTTAGGGTCAGGGTTAGGGTTAGGGTCAGG...

1942515 rows × 14 columns

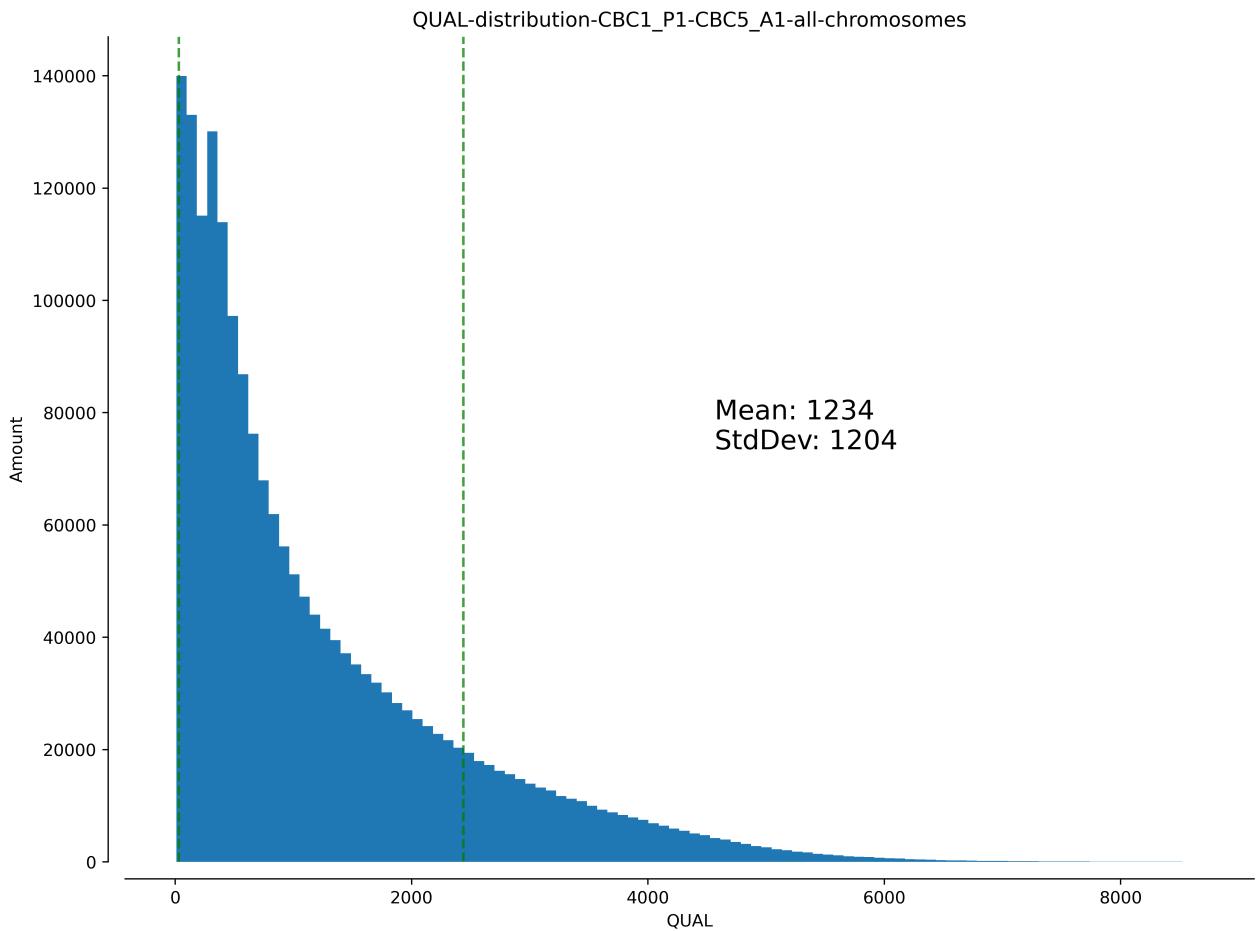


Verify DP Histogram Cutoff Off by Mean±2StdDev

In [30]: `plot_variant_hist(samples, vcf_df_01, 'all', 'DP', bins=200, xmax=200)`



In [31]: `plot_variant_hist(samples, vcf_df_01, 'all', 'QUAL', bins=100, MSTD=True)`



Contingency Table After DP Cutoff by Mean±2StdDev

```
In [32]: contingency_table_2 = contingency_table(samples, vcf_df_01, 'all')
```

Contingency Table - Chromosome all

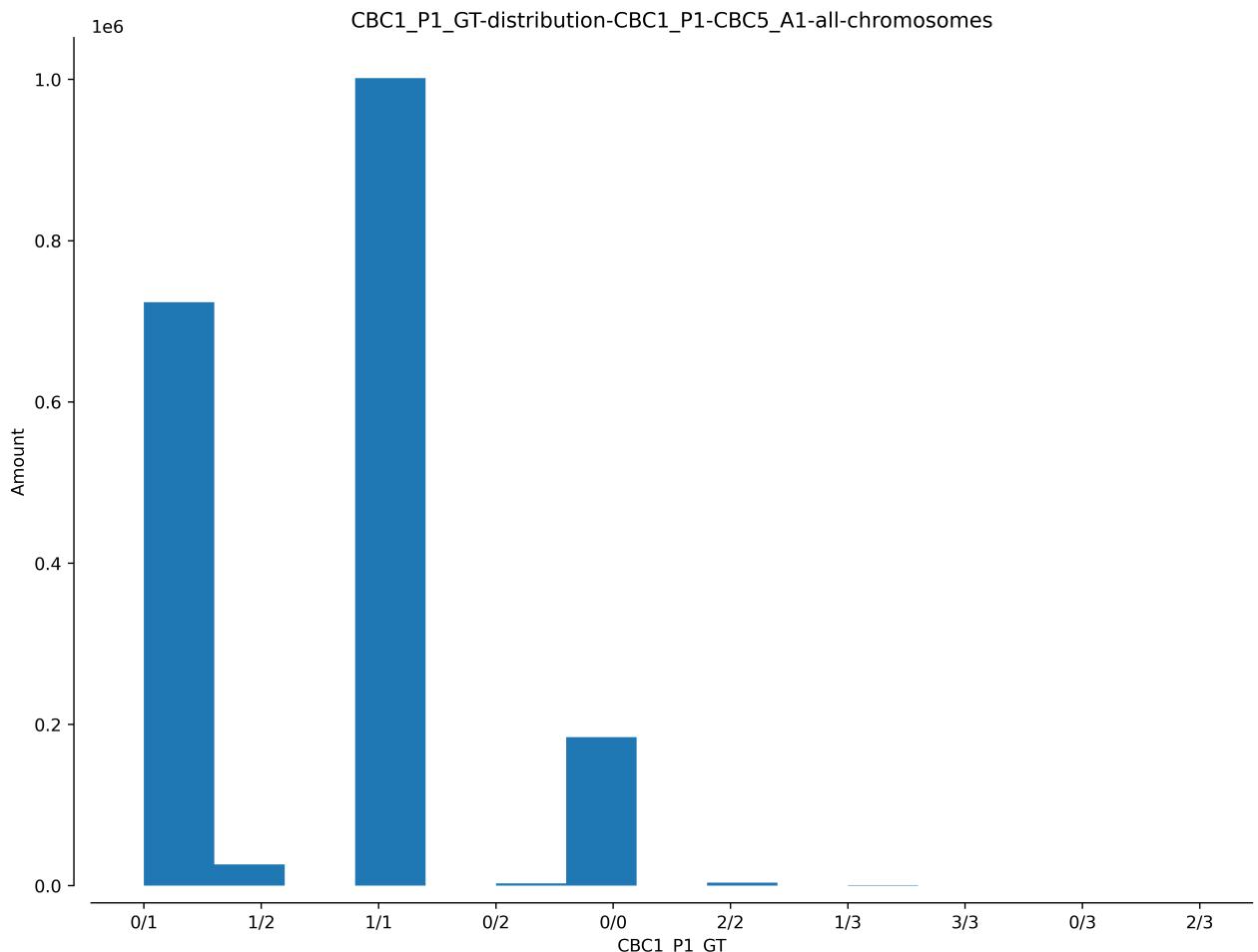
		CBC5_A1_GT			
		0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	51039	132478	43507
	0/1	284832	255445	178544	43507
	1/1	211386	19531	765753	43507
	other	43507	43507	43507	43507

GT Plot After DP Cutoff by Mean±2StdDev

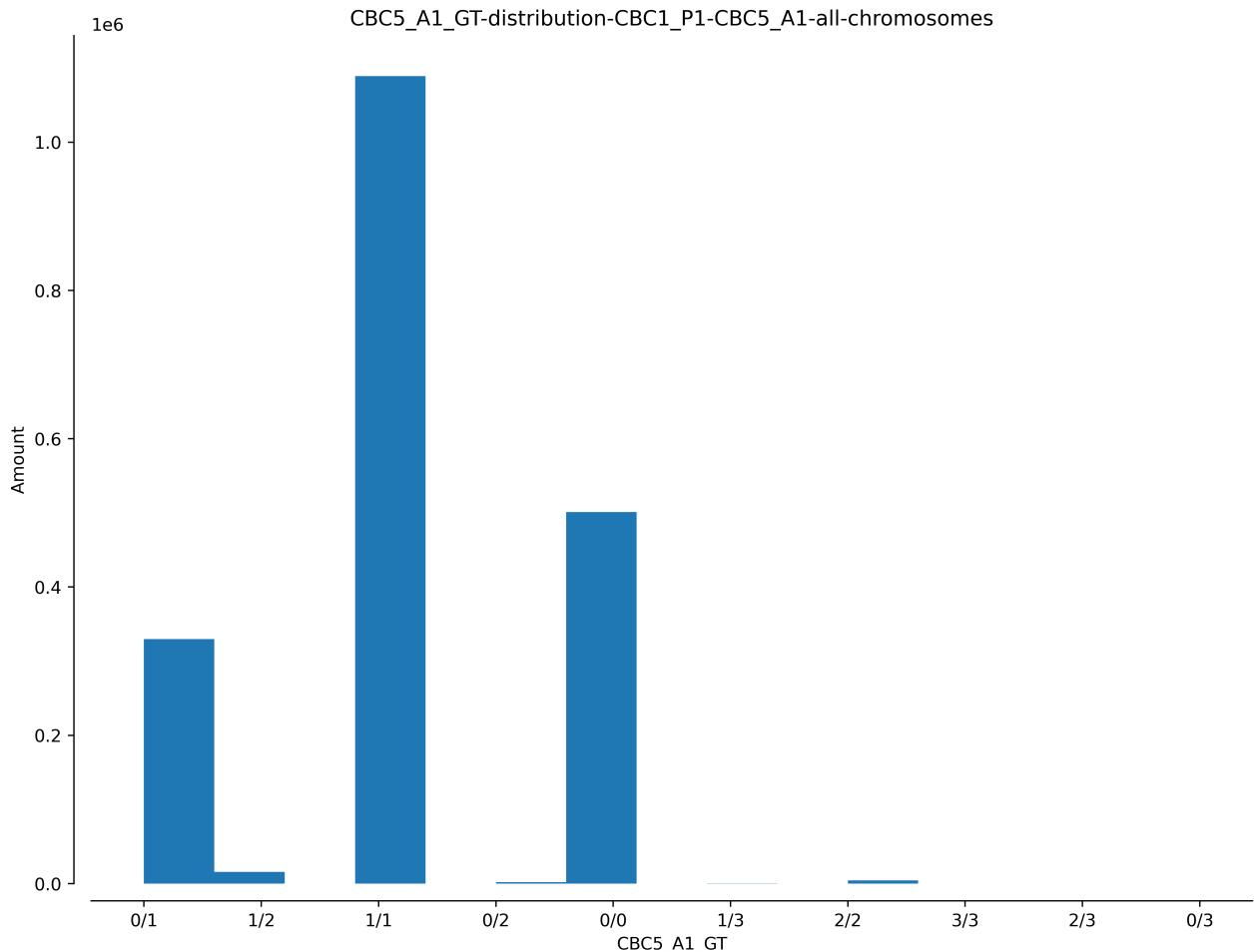
```
In [33]: # plt.close('all')
# GTplot(samples, vcf_df_01, chrom_len_00)
```

Histogram 'GT' Attribute after DP Cutoff

```
In [34]: plot_variant_hist(samples, vcf_df_01, 'all', '%s_GT' % progenitor, bins=15)
```



```
In [35]: plot_variant_hist(samples, vcf_df_01, 'all', '%s_GT' % mutant, bins=15)
```



PART 3: Extract *snp* from *TYPE* Attribute

Extract *snp* *TYPE* Attribute

In [36]:

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

Out[36]:

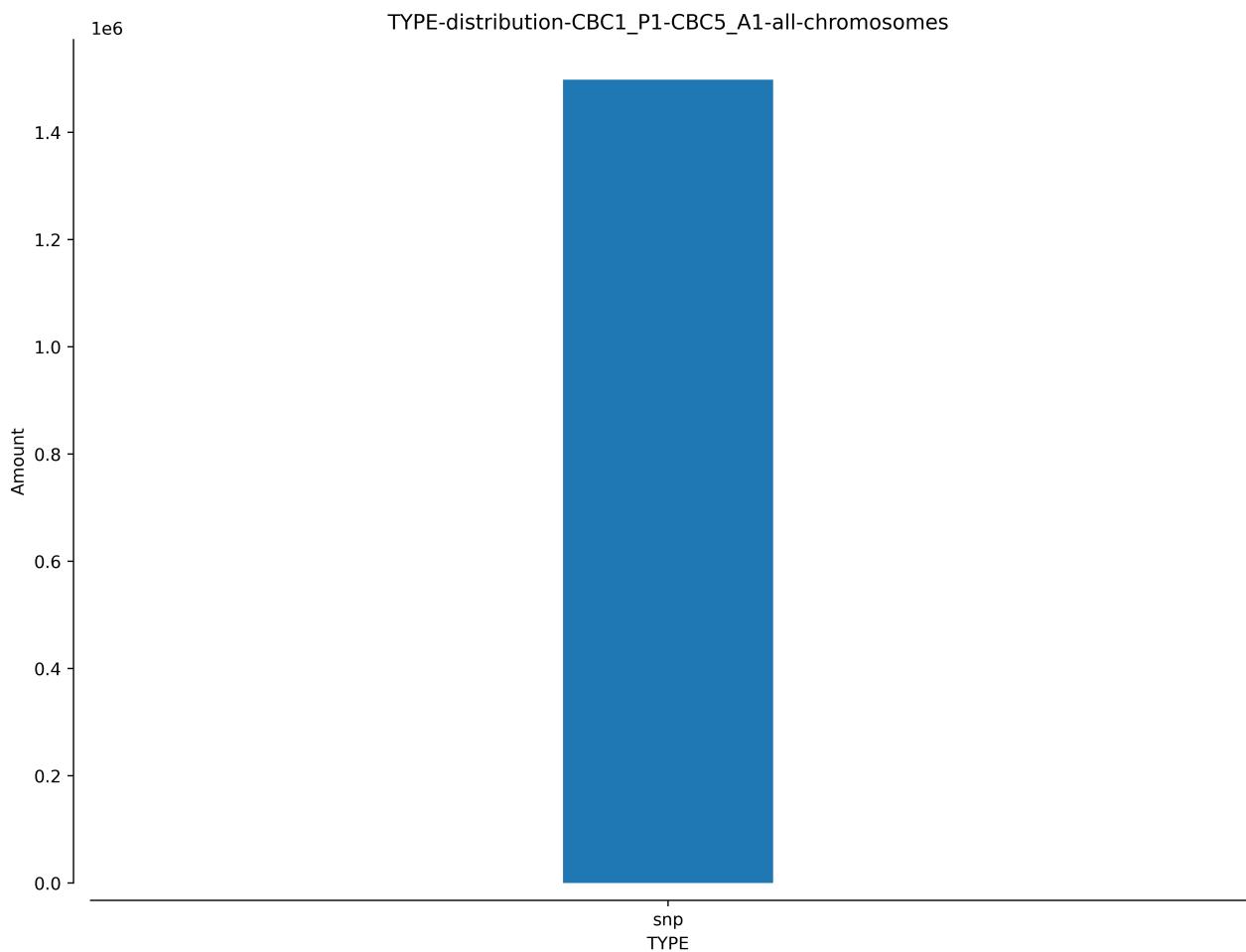
	CHROM	POS	REF	ALT	QUAL	DP	CBC1_P1_DP	CBC5_A1_DP	CBC1_P
0	NC_040279.1	912	G	T	571.966003	121		42	79
1	NC_040279.1	1173	C	T	321.862000	57		34	23
2	NC_040279.1	1390	C	T	214.576996	46		25	21
3	NC_040279.1	1424	T	C	420.058014	71		35	36
4	NC_040279.1	1699	C	T	2160.239990	125		52	73
...
1498055	NC_040289.1	41658237	C	T	384.597992	141		70	71
1498056	NC_040289.1	41658862	G	A	94.799599	128		70	58
1498057	NC_040289.1	41659114	T	C	52.265900	30		17	13
1498058	NC_040289.1	41659137	G	C	35.938702	41		24	17

```
CHROM      POS   REF   ALT    QUAL   DP  CBC1_P1_DP  CBC5_A1_DP  CBC1_P
1498059  NC_040289.1  41667148    T     C    158.554993  14       10        4
```

1498060 rows × 14 columns

TYPE.snp Histogram Verification

```
In [37]: plot_variant_hist(samples, vcf_df_02, 'all', 'TYPE', bins=5)
```



Contingency Table -.snp TYPE only

```
In [38]: contingency_table_3 = contingency_table(samples, vcf_df_02, 'all')
```

Contingency Table - Chromosome all

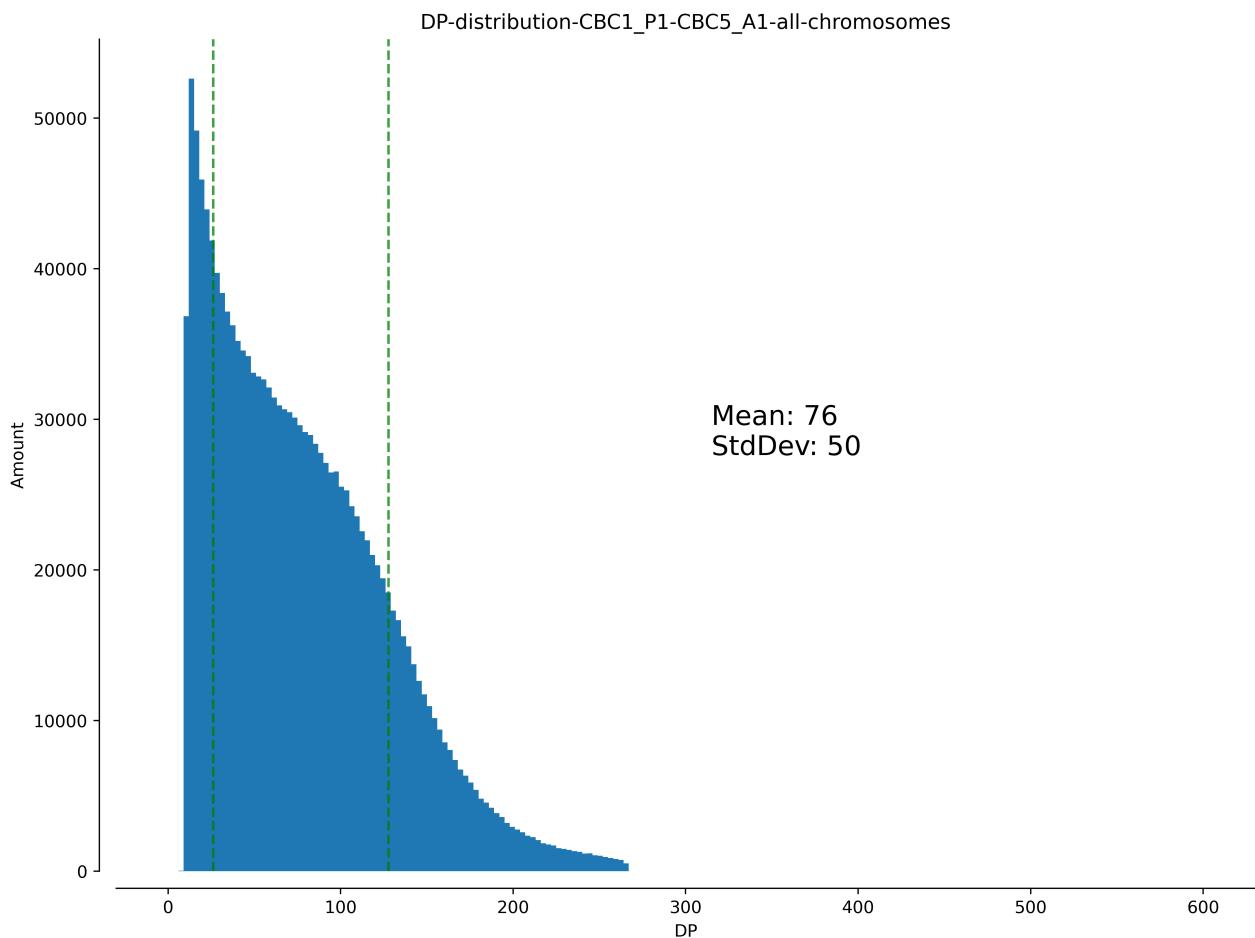
		CBC5_A1_GT			
		0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	40040	105589	8593
	0/1	230637	202120	142758	8593
	1/1	169172	14918	584233	8593
	other	8593	8593	8593	8593

GT Plot -.snp TYPE only

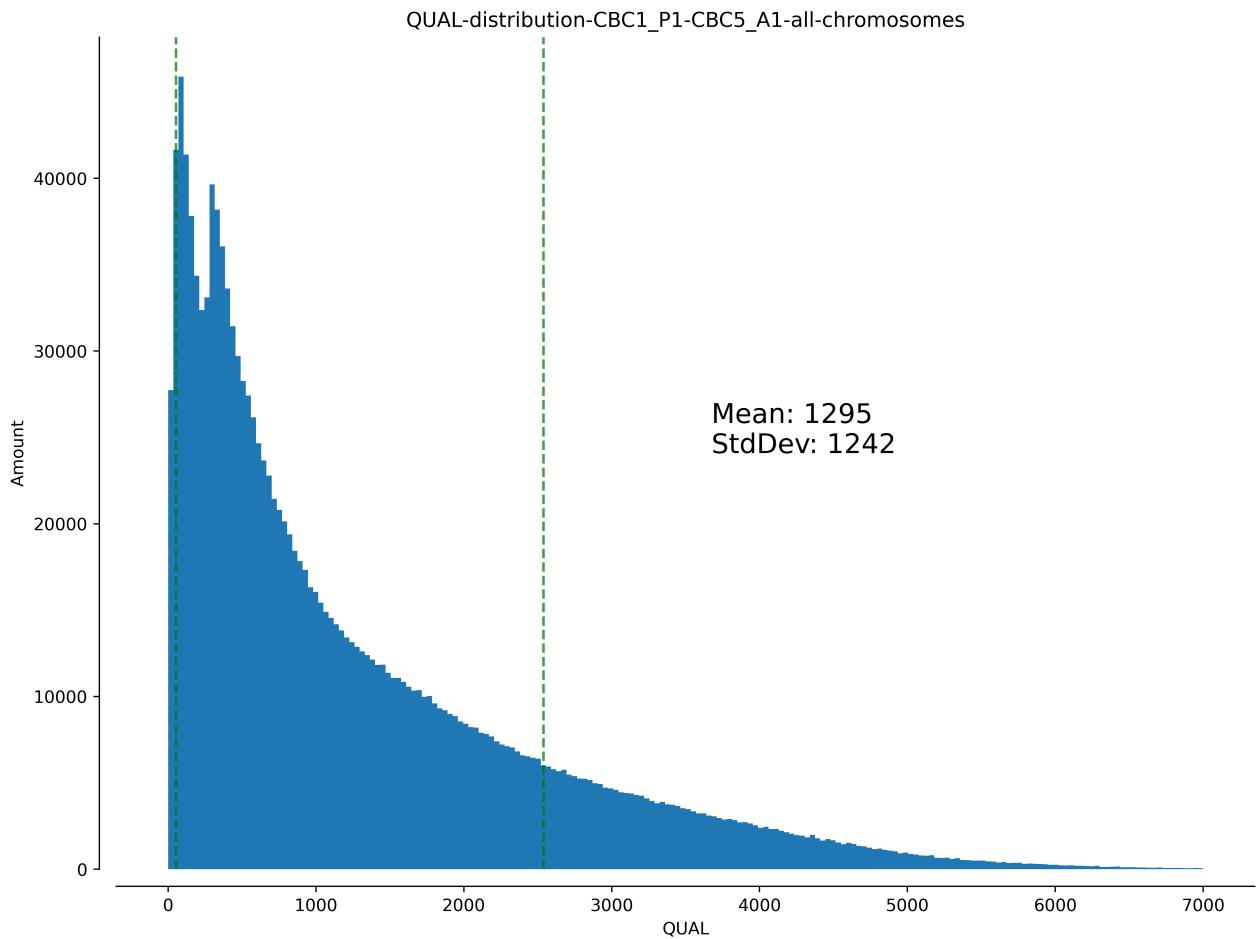
```
In [39]: # plt.close('all')
# GTplot(samples, vcf_df_02, chrom_len_00)
```

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

```
In [40]: plot_variant_hist(samples, vcf_df_02, 'all', 'DP', bins=200, MSTD=True, xmax=600)
```

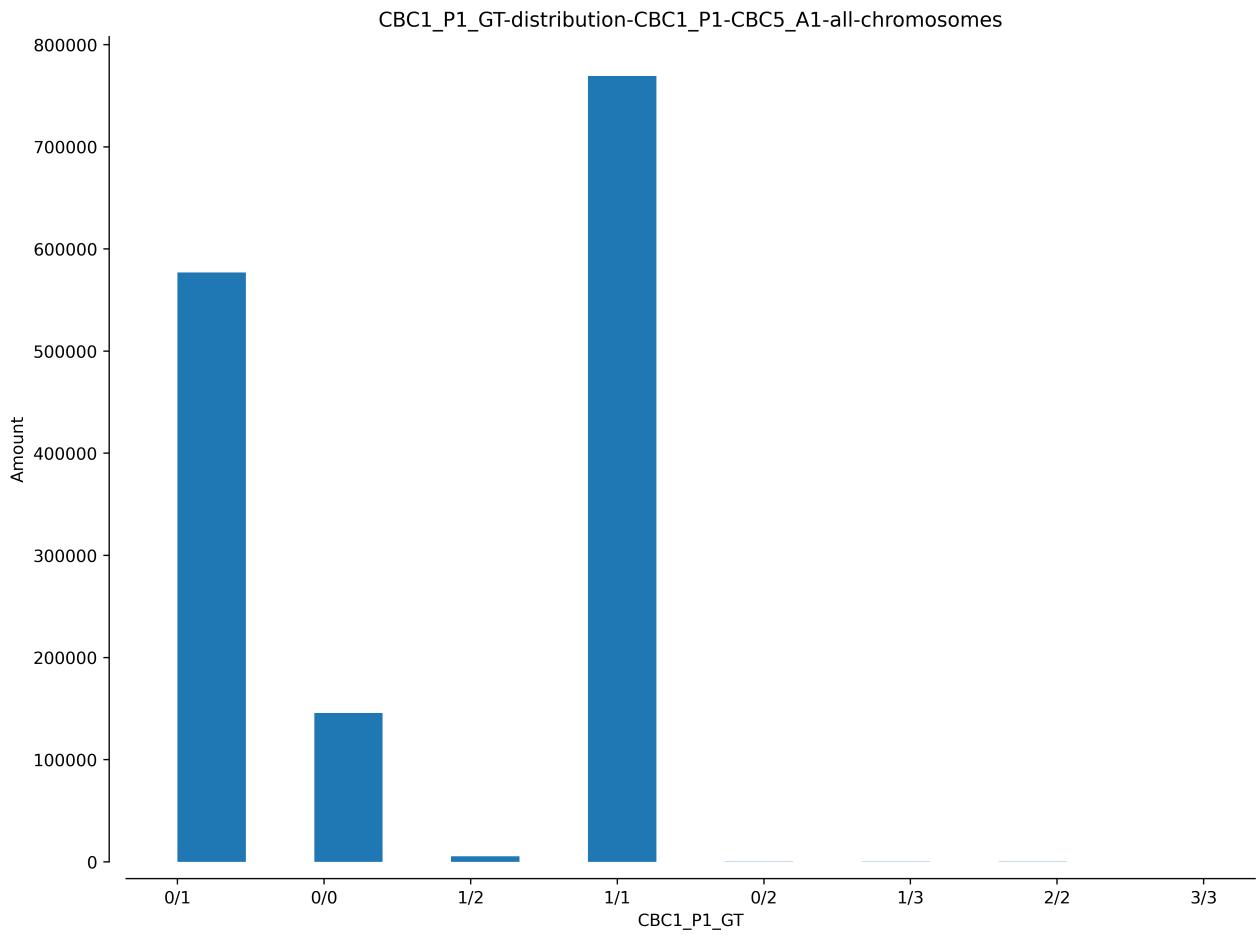


```
In [41]: plot_variant_hist(samples, vcf_df_02, 'all', 'QUAL', bins=200, MSTD=True, xmax=600)
```

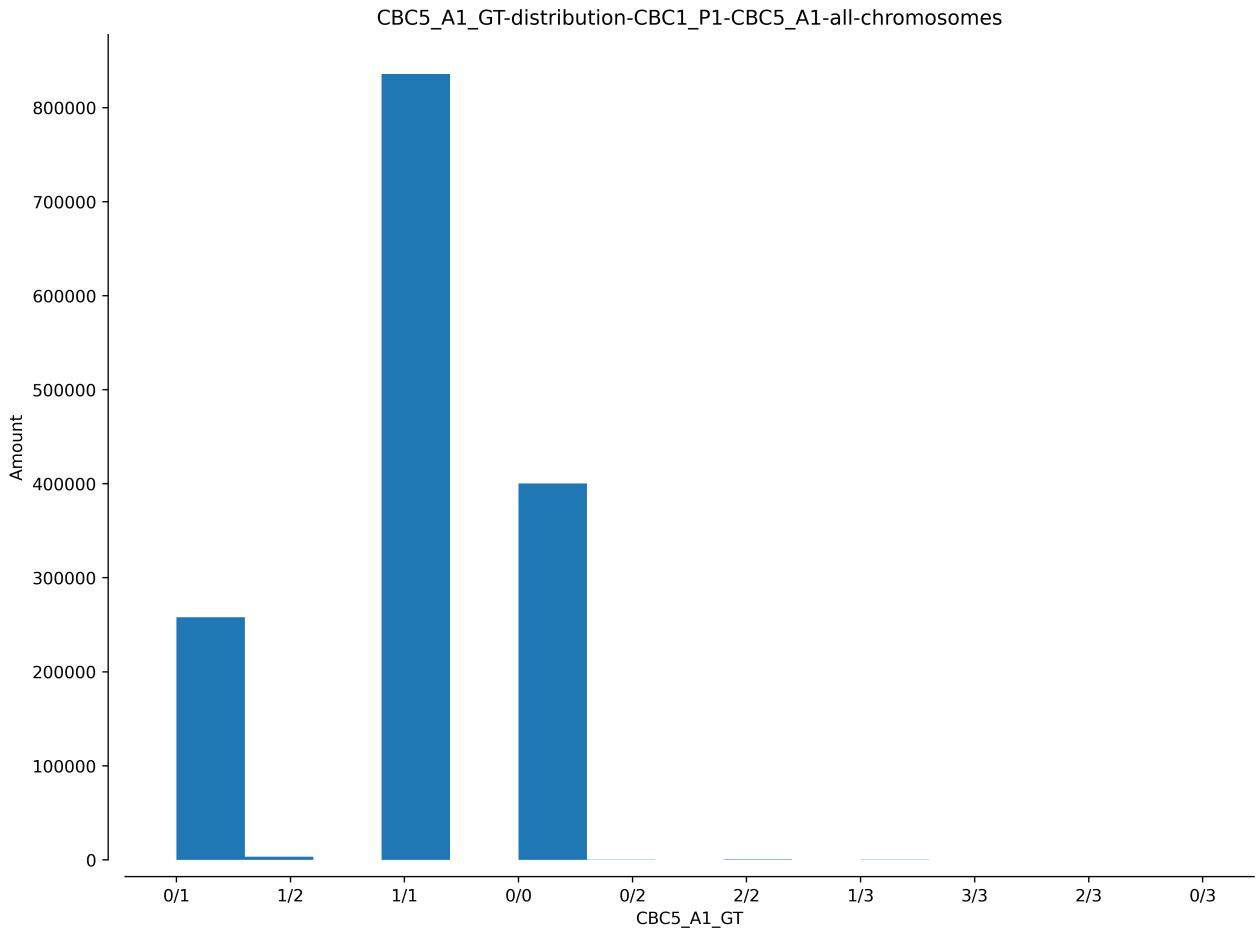


In [42]:

```
plot_variant_hist(samples, vcf_df_02, 'all', '%s_GT' % progenitor, bins=15)
```



```
In [43]: plot_variant_hist(samples, vcf_df_02, 'all', '%s_GT' % mutant, bins=15)
```



PART 4: Cutting Off by Mean \pm StdDev Histograms of *QUAL* Attribute

```
In [44]: # 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 \pm StdDev

```
In [45]: # plot_variant_hist(samples, vcf_df_03, 'all', 'DP', bins=200, xmax=200)
```

```
In [46]: # plot_variant_hist(samples, vcf_df_03, 'all', 'QUAL', bins=100, xmax=3500)
```

Contingency Table After QUAL Cutoff by Mean \pm StdDev

```
In [47]: # contingency_table_4 = contingency_table(samples, vcf_df_03, 'all')
```

GT Plot After QUAL Cutoff by Mean±StdDev

```
In [48]: # plt.close('all')
# GTplot(samples, vcf_df_03, chrom_len_00)
```

Histograms after QUAL Cutoff by Mean±StdDev

```
In [49]: # plot_variant_hist(samples, vcf_df_03, 'all', 'PAHAT_1_GT', bins=9)
```

```
In [50]: # 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 [51]: progenitor_gts_filter = "CBC1_P1_GT != ./., CBC1_P1_GT != 0/2, CBC1_P1_GT != 1/2"
vcf_df_04 = filter_vcf(vcf_df_02, progenitor_gts_filter)

mutant_gts_filter = "CBC5_A1_GT != ./., CBC5_A1_GT != 0/2, CBC5_A1_GT != 1/2, CE
vcf_df_04 = filter_vcf(vcf_df_04, mutant_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[51]:
```

	CHROM	POS	REF	ALT	QUAL	DP	CBC1_P1_DP	CBC5_A1_DP	CBC1_P1
0	NC_040279.1	912	G	T	571.966003	121	42	79	
1	NC_040279.1	1173	C	T	321.862000	57	34	23	
2	NC_040279.1	1390	C	T	214.576996	46	25	21	
3	NC_040279.1	1424	T	C	420.058014	71	35	36	
4	NC_040279.1	1699	C	T	2160.239990	125	52	73	
...
905229	NC_040289.1	41658237	C	T	384.597992	141	70	71	
905230	NC_040289.1	41658862	G	A	94.799599	128	70	58	
905231	NC_040289.1	41659114	T	C	52.265900	30	17	13	
905232	NC_040289.1	41659137	G	C	35.938702	41	24	17	
905233	NC_040289.1	41667148	T	C	158.554993	14	10	4	

905234 rows × 14 columns

Contingency Table after GT Filtering

```
In [52]: contingency_table_5 = contingency_table(samples, vcf_df_04, 'all')
```

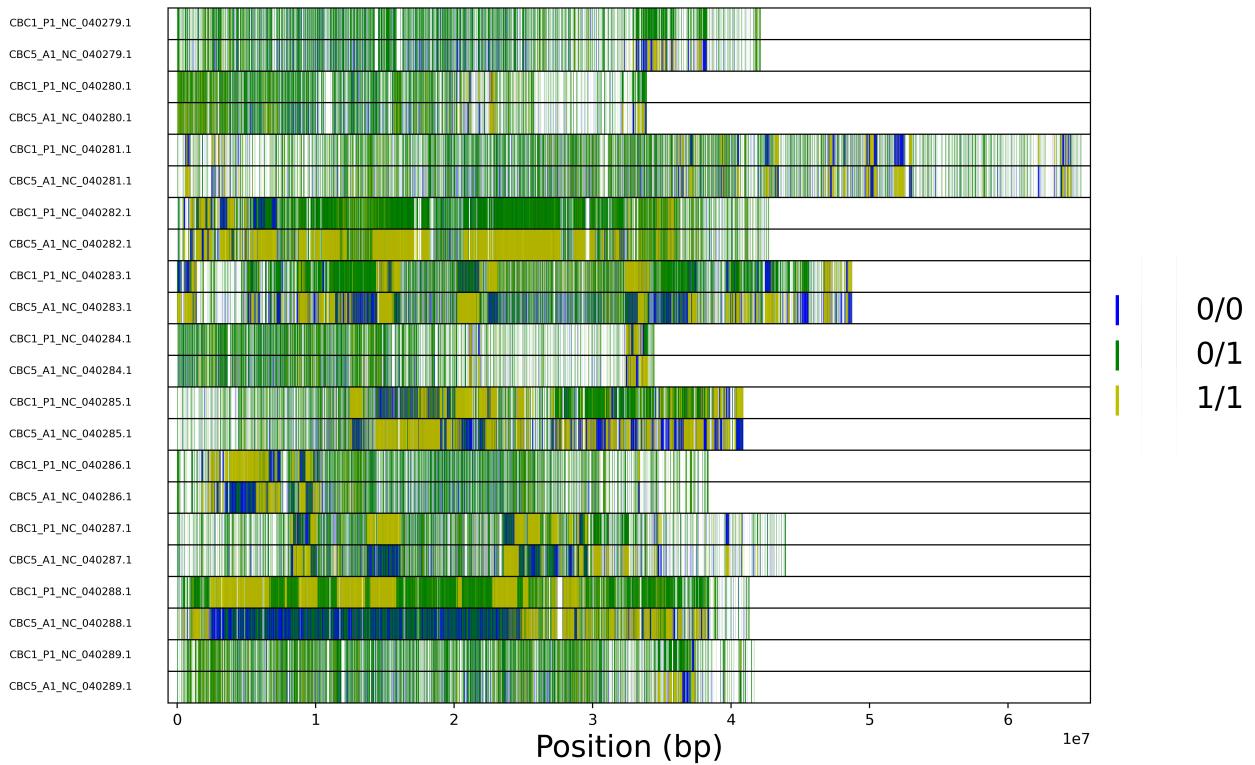
Contingency Table - Chromosome all

		CBC5_A1_GT			
		0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	40040	105589	0
	0/1	230637	202120	142758	0
	1/1	169172	14918	0	0
	other	0	0	0	0

GT Plot after GT Filtering

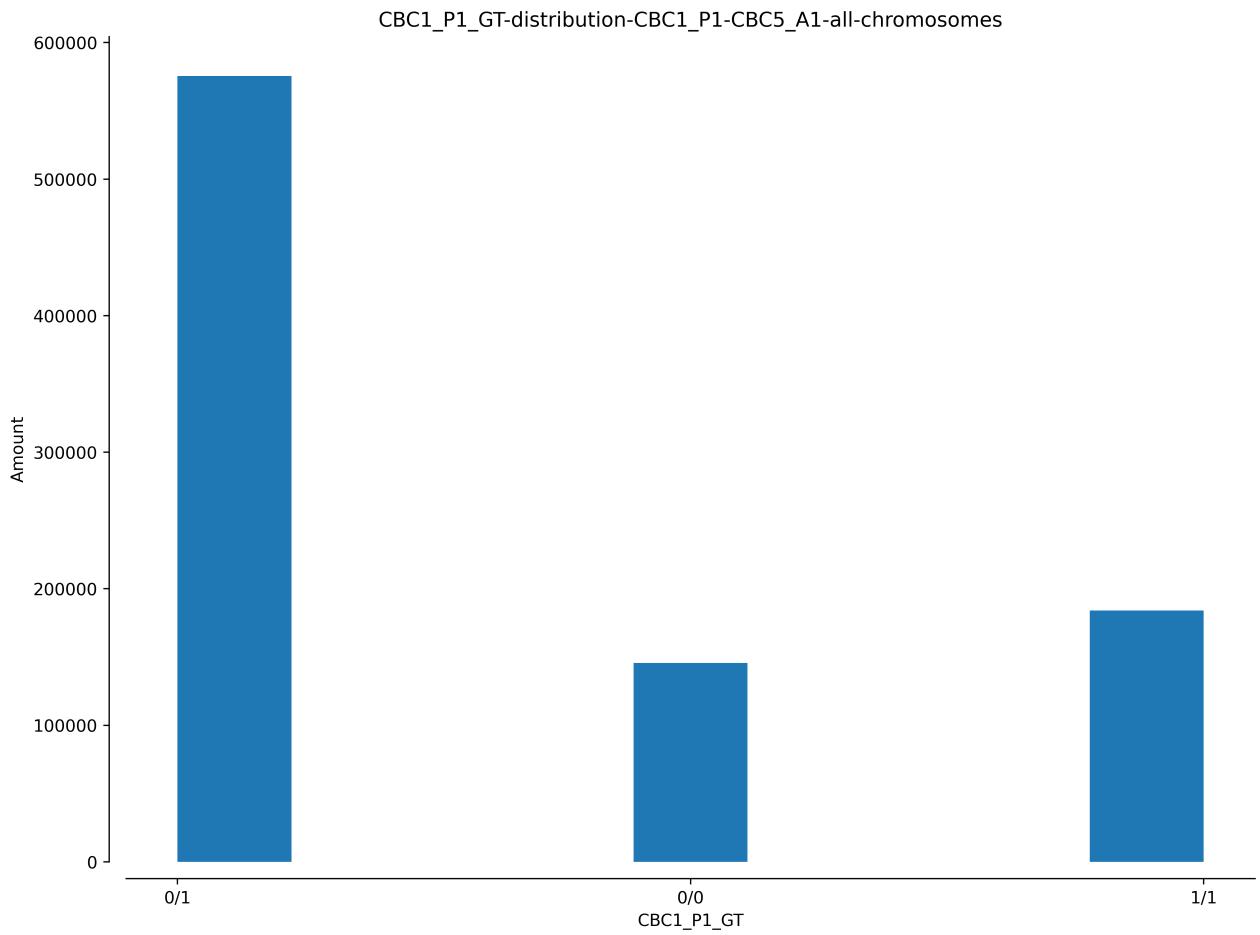
```
In [53]: plt.close('all')
GTplot(samples, vcf_df_04, chrom_len_00)
```

gt-plot-CBC1_P1-CBC5_A1

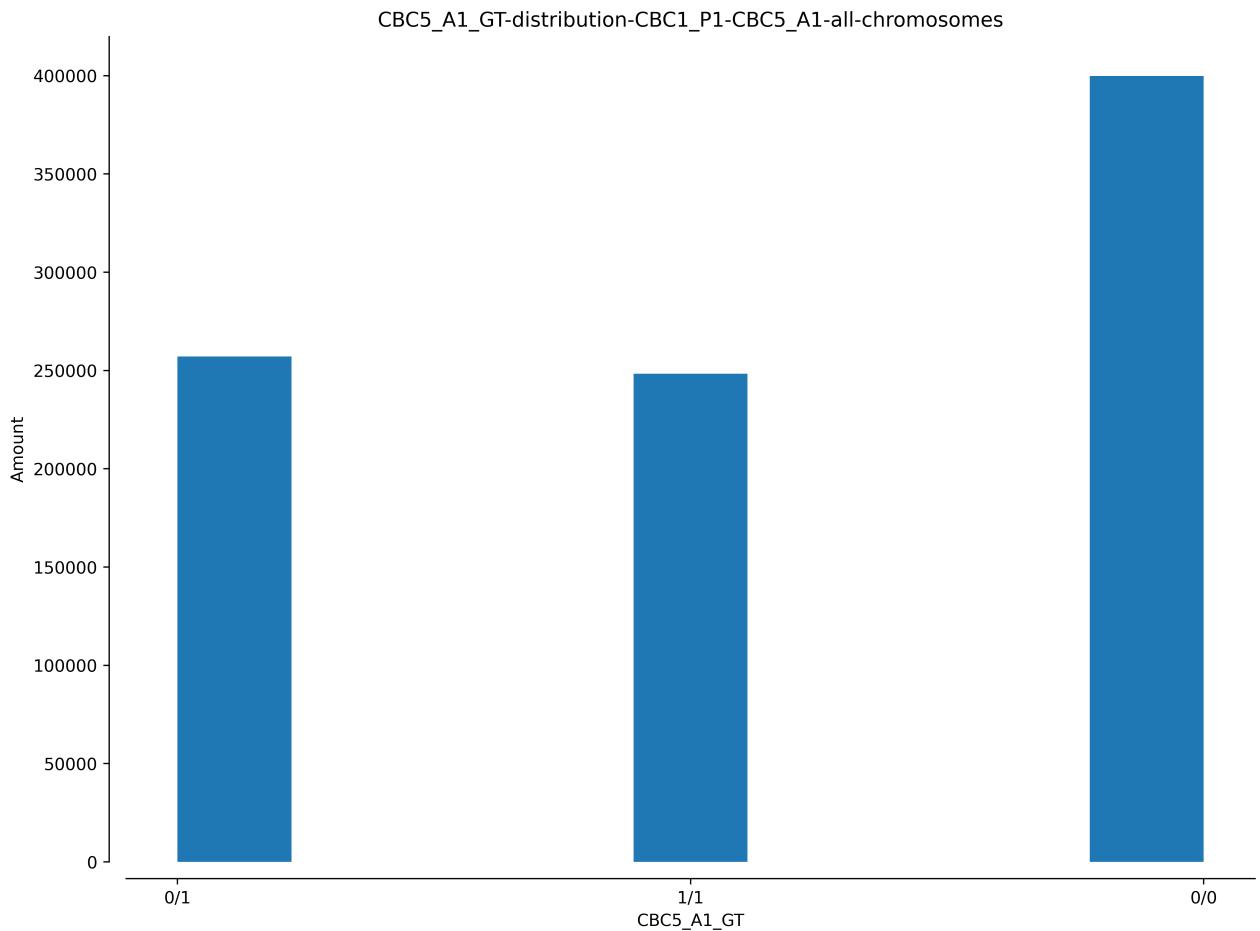


Histograms GT after GT Filtering

```
In [54]: plot_variant_hist(samples, vcf_df_04, 'all', '%s_GT' % progenitor, bins=9)
```



```
In [55]: plot_variant_hist(samples, vcf_df_04, 'all', '%s_GT' % mutant, bins=9)
```



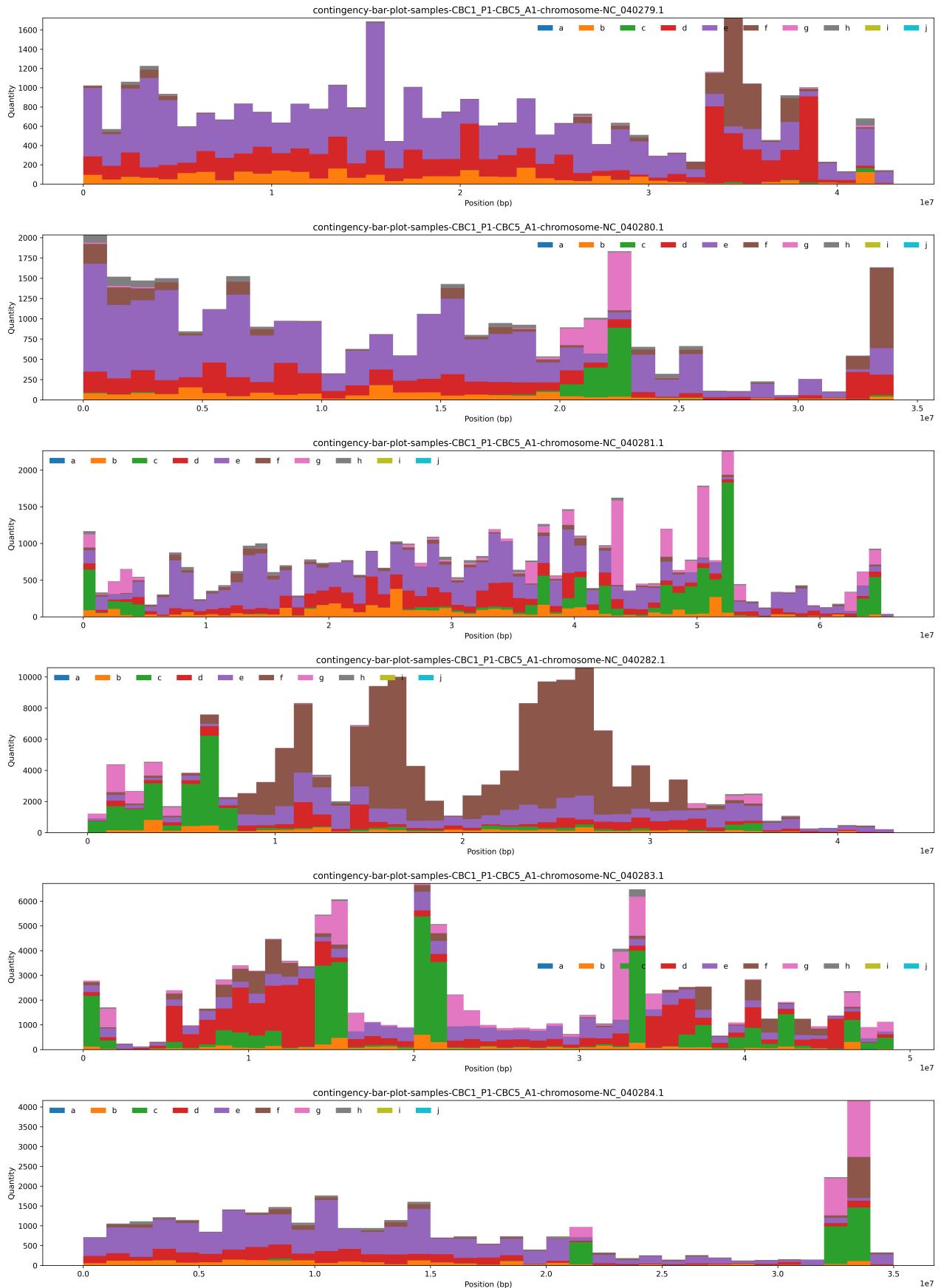
PART 6: Stacked Bar Plots

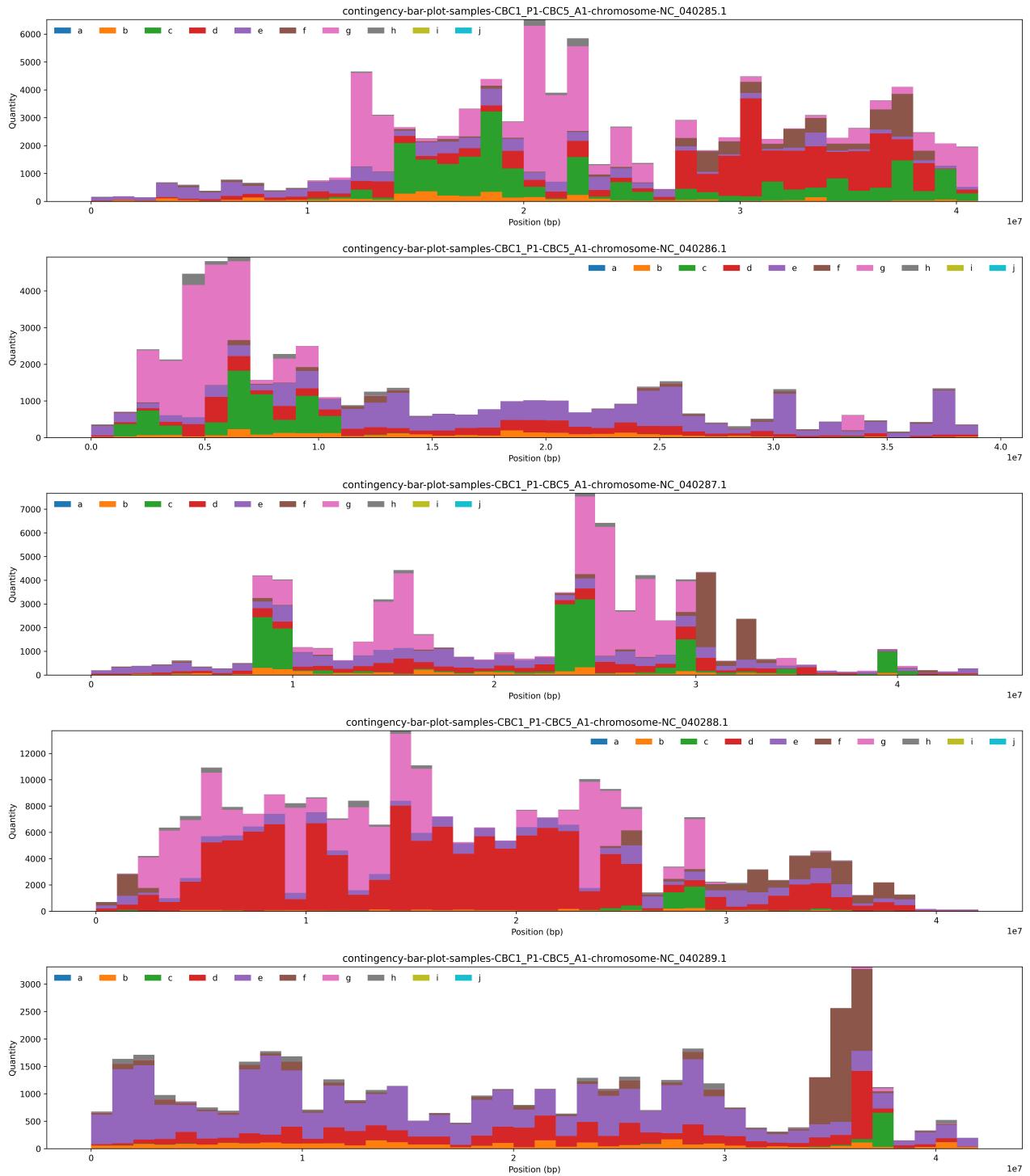
In [56]: `ct_guide()`

Out[56]:

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



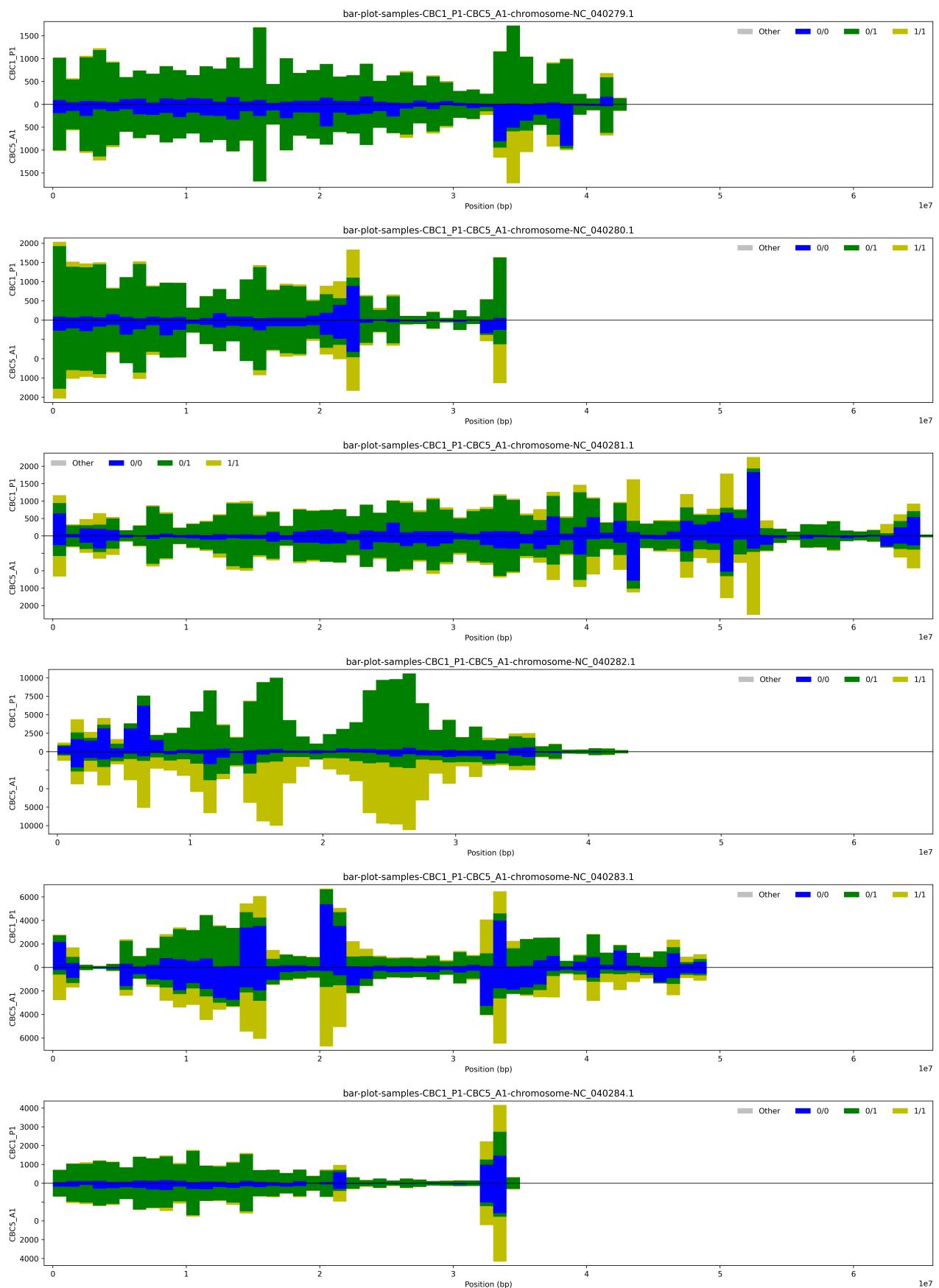


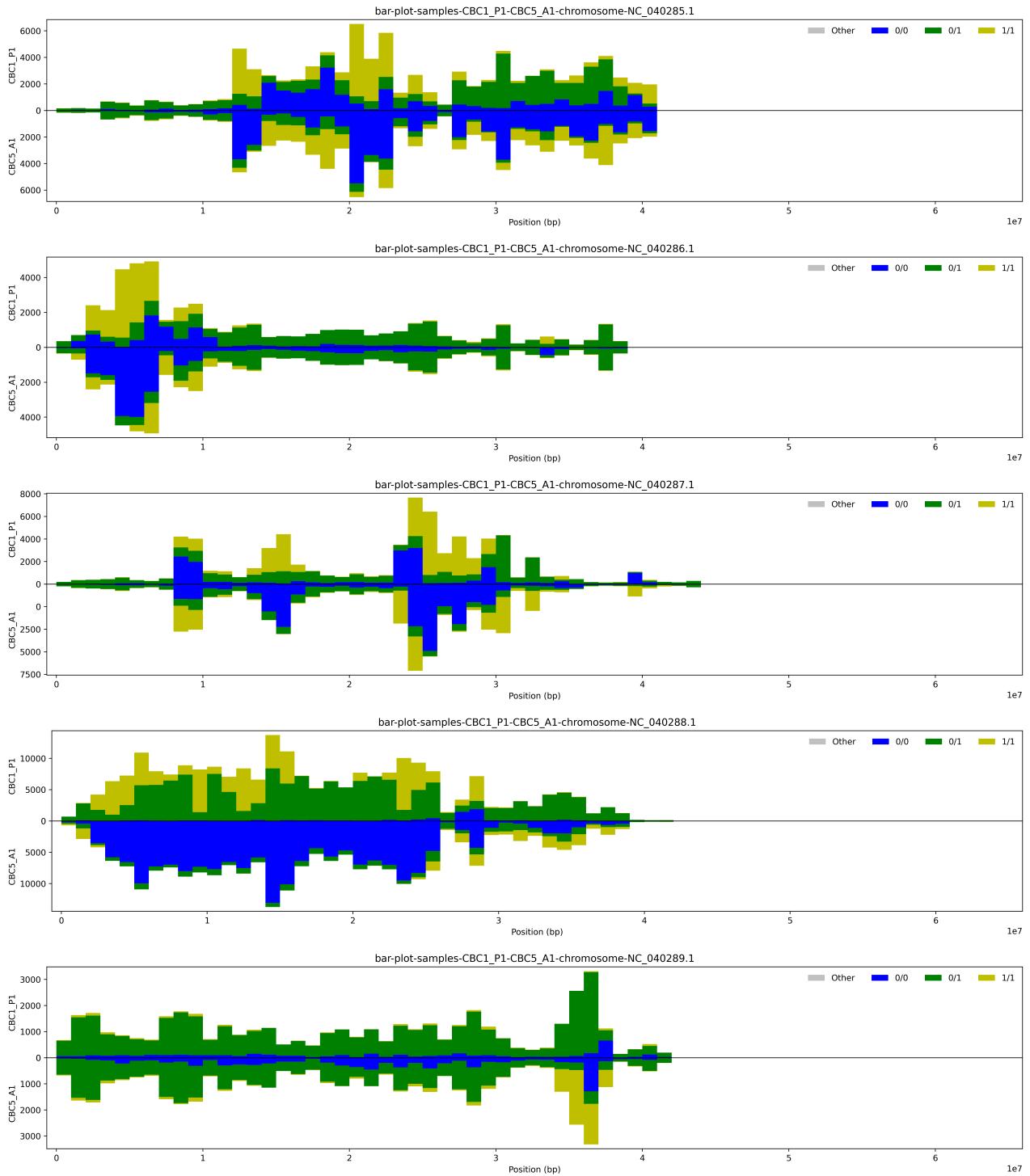
PART 7: Bar Plots per Chromosome

In [58]:

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

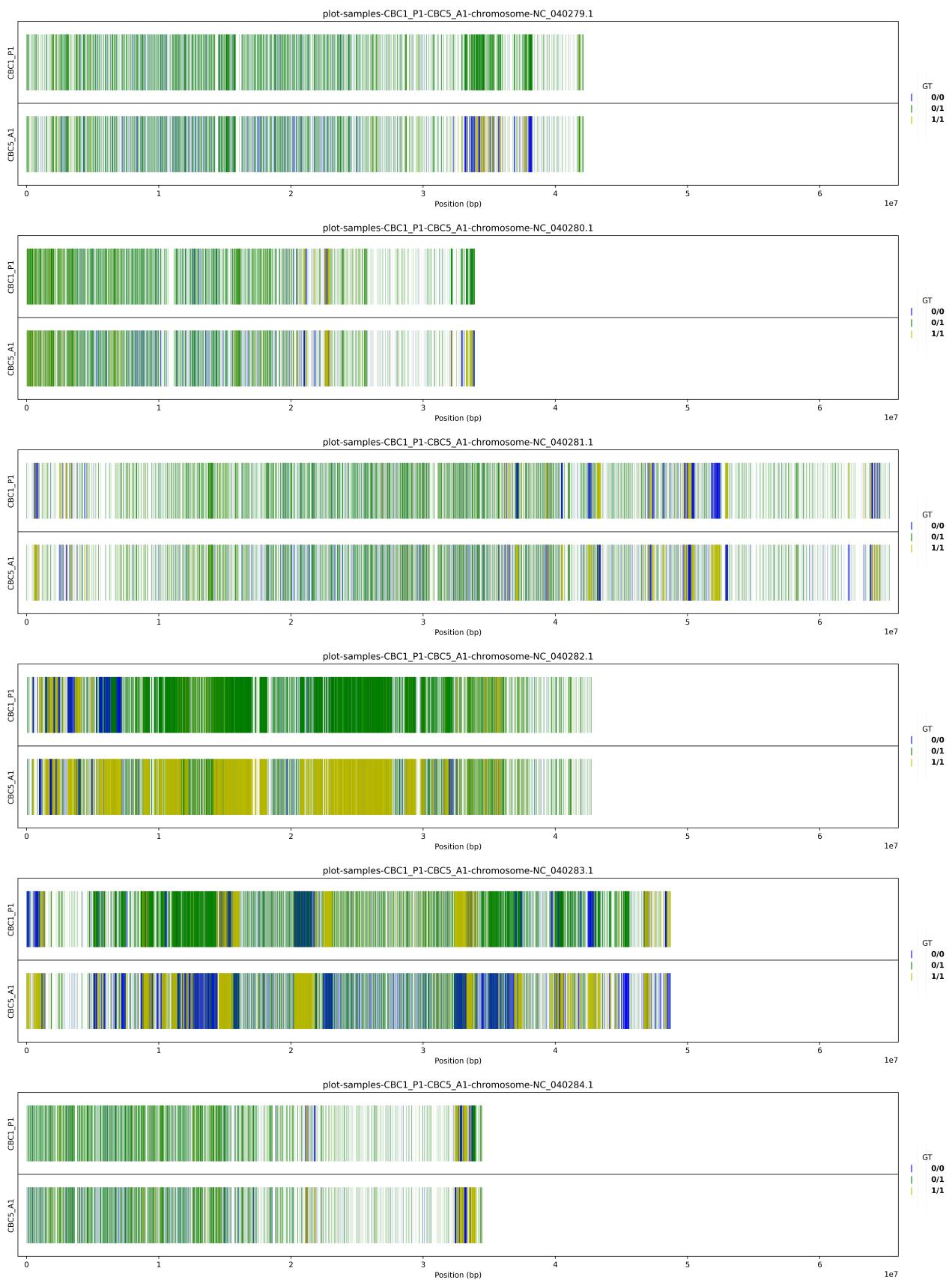
plt.close('all')
GTbarPlots(samples, vcf_df_04, chrom_len_00, window_size)
```

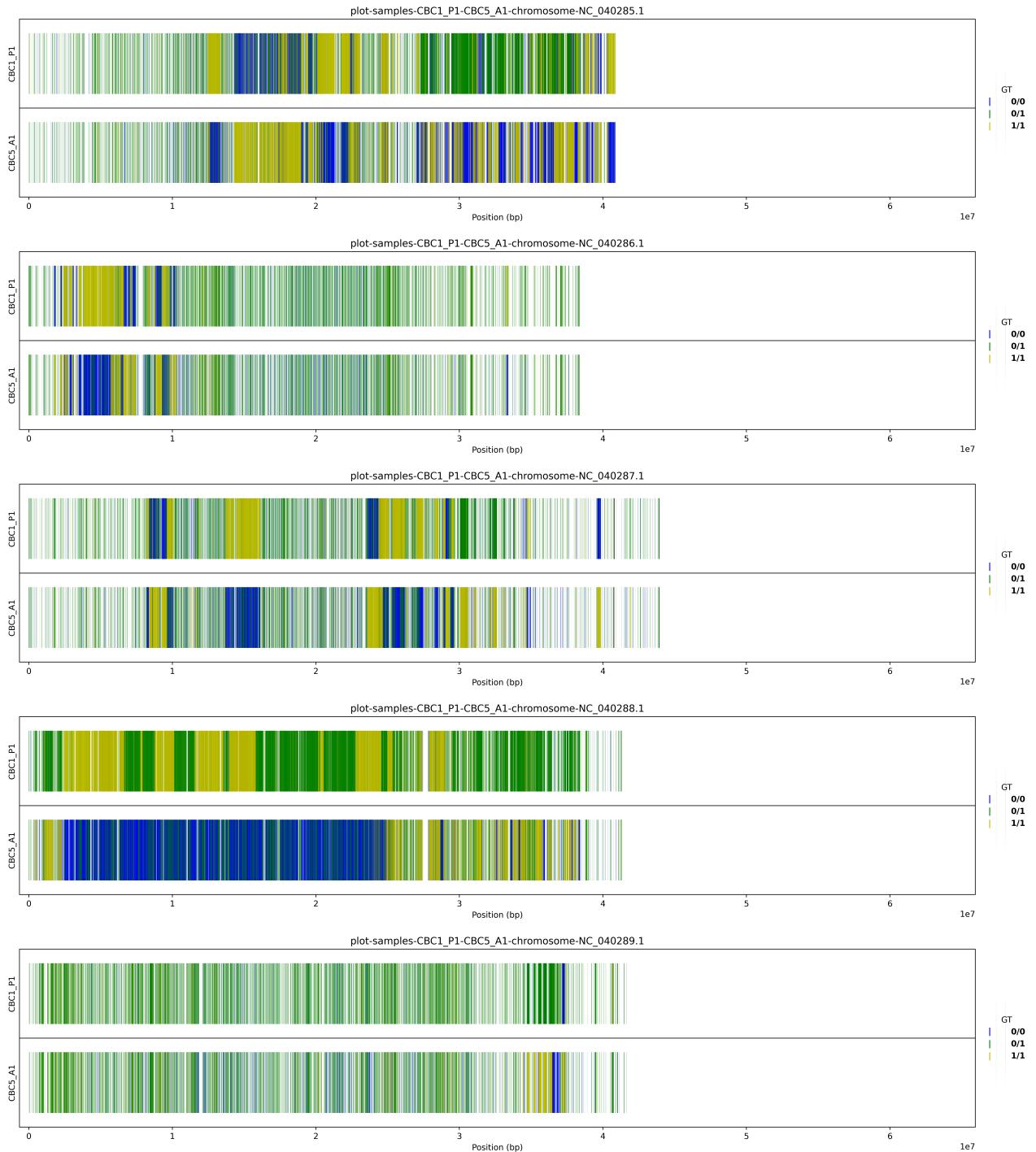




PART 8: GT Plots per Chromosome

```
In [59]: plt.close('all')
GTplots(samples, vcf_df_04, chrom_len_00)
```





PART 9: Contingency Table per Chromosome

In [60]:

```
import dataframe_image as dfi

for chromosome in chrom_len_00.index:
    chromosome_df = vcf_df_04[ vcdf_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_040279.1

CBC5_A1_GT

	0/0	0/1	1/1	other	
CBC1_P1_GT	0/0	0	2849	100	0
	0/1	9173	16676	2646	0
	1/1	61	418	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040280.1

	CBC5_A1_GT	0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	1884	1466	0
	0/1	5160	15796	2694	0
	1/1	1443	884	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040281.1

	CBC5_A1_GT	0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	4400	7414	0
	0/1	7413	18737	1827	0
	1/1	5653	1050	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040282.1

	CBC5_A1_GT	0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	6971	19650	0
	0/1	14539	30212	88640	0
	1/1	5755	1208	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040283.1

	CBC5_A1_GT	0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	5380	29771	0
	0/1	29519	17174	9243	0
	1/1	14442	985	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040284.1

	CBC5_A1_GT	0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	2203	2942	0
	0/1	4892	15518	2307	0
	1/1	2664	818	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040285.1

	CBC5_A1_GT	0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	3706	19924	0
	0/1	22736	10967	7241	0
	1/1	26833	1185	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040286.1

	CBC5_A1_GT	0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	2762	6205	0

0/1	6624	17047	1278	0
1/1	13794	1374	0	0
other	0	0	0	0

Contingency Table - Chromosome NC_040287.1

		CBC5_A1_GT			
		0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	3667	13460	0
	0/1	8492	13585	6454	0
	1/1	26026	1206	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040288.1

		CBC5_A1_GT			
		0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	3094	3879	0
	0/1	113815	22685	14159	0
	1/1	72334	4425	0	0
	other	0	0	0	0

Contingency Table - Chromosome NC_040289.1

		CBC5_A1_GT			
		0/0	0/1	1/1	other
CBC1_P1_GT	0/0	0	3124	778	0
	0/1	8274	23723	6269	0
	1/1	167	1365	0	0
	other	0	0	0	0