Preprocessing: input is a variant caller file, interested in the unique ID: columns
#CHROM POS REF ALT

Also interested in AF (allele frequency)

The database will have ONE singular master file that has 4 columns being the unique ID, and every column after will be 0 or 1, based on the column title designating the samples ID, 0 means that sample doesnt have the variant 1 means it does USE FEATHER FORMAT

Simple to extract uqid, harder to get af because its varied where it could be

Database DataFrame

	A	B	C
15CG	1	0	0
17CG	0	1	0
18CG	0	0	1

Input DataFrame

D
17CG
16CG

Expected Final DataFrame

	A	B	C	D
15CG	1	0	0	0
16CG	0	0	0	1
17CG	0	1	0	1
18CG	0	0	1	0

Updated Input DataFrame

	A	B	C	D
17CG	0	0	0	1
16CG	0	0	0	1

Copy DataFrame

	A	B	C	D
15CG	1	0	0	1
17CG	0	1	0	1
18CG	0	0	1	1
17CG	0	0	0	1
16CG	0	0	0	1

Step 1:
Add input DataFrame rows (filled with 0s) to a copy of database DataFrame.
Add column D (filled with 1s)

Copy DataFrame

	A	B	C	D
15CG	1	0	0	1
17CG	0	1	0	1
18CG	0	0	1	1
17CG	0	0	0	1
16CG	0	0	0	1

Step 2:
Extract first occurrence of duplicate rows

*after this step, copy dataframe no longer needed

First Occurence Duplicates DataFrame

	A	B	C	D
17CG	0	1	0	1

Database DataFrame

	A	B	C	D
15CG	1	0	0	0
17CG	0	1	0	0
18CG	0	0	1	0

Updated Input DataFrame

	A	B	C	D
17CG	0	0	0	1
16CG	0	0	0	1

First Occurence Duplicates DataFrame

	A	B	C	D
17CG	0	1	0	1

Updated Database DataFrame

	A	B	C	D
15CG	1	0	0	0
17CG	0	1	0	0
18CG	0	0	1	0
17CG	0	0	0	1
16CG	0	0	0	1
17CG	0	1	0	1

Step 3:

Add column D (filled with 0s) to database DataFrame. Add updated input DataFrame and first occurrence duplicates DataFrame to bottom of database DataFrame

Updated Database DataFrame

	A	B	C	D
15CG	1	0	0	0
17CG	0	1	0	0
18CG	0	0	1	0
17CG	0	0	0	1
16CG	0	0	0	1
17CG	0	1	0	1

Step 5:
Remove duplicate rows of updated database dataframe, keep last occurrence

Final Database DataFrame

	A	B	C	D
15CG	1	0	0	0
16CG	0	0	0	1
17CG	0	1	0	1
18CG	0	0	1	0

Step 6:
Sort final database DataFrame
Sort full input DataFrame
Save both to database storage

Sorted Input DataFrame

D
17CG
16CG

→

	FILTER	DP	AF	XYZ
16CG	EXONIC	23	.14	123
17CG	EXONIC	64	.19	456