The purpose of the genotype submission guide is to facilitate the submission of the [bulk of the study’s genotype data](#bulk). It describes what types of information are anticipated in the submission and how the submitted data will be processed and distributed.

1. **Preparing genotypes for submission to dbGaP**
2. **Study level information**
3. ***Individual info.*** Please provide information about the study and control subjects and samples in a standard form described in the dbGaP File Submission Guide.Please note that only genotyped samples to be released (including duplicated and control samples) should be listed in the dbGaP Subject Sample Mapping Data File. At the same time, every study or control subject mentioned in any study dataset file must be listed in the dbGaP Subject Consent Data File with corresponding consent information. Individual level genotypes will be verified, packed, and distributed according to the information provided in the Subject Consent, Subject Sample Mapping, and Pedigree files.
4. ***Marker info.*** Please provide information on the marker set used for the genotyping.
5. ***Quality Control (QC).*** Please submit the results of your genotyping QC analysis and metrics, if available. Preliminary results of GWA may be submitted as a pre-computed component intended to validate the quality of biologically-relevant variable(s) and the consistency between phenotype and genotype id spaces. If the quality of genotyped samples is known, please provide a list of subject/sample IDs, whose submitted genotypes are in low quality, and unacceptable to analysis.
6. **Individual level genotype data**
7. ***Data types and formats*.** Although the majority of dbGaP studies today are SNP-based GWA studies, dbGaP accepts genotypes for any quantity, any technology-based markers and genes.

Files of different data types and platforms have to be packed separately with clear descriptions of their contents and column headers. The submitter also needs to provide sample-file mapping info for each package.

To ensure the comprehensiveness of a study’s representation, dbGaP encourages investigators to submit raw and normalized data derived at each step of data acquisition and transformation. Such data representation provides the flexibility to validate, re-evaluate, and re-organize datasets in accordance with one’s views and needs.

Genotypes can be submitted to dbGaP in a variety of formats:

1. Raw level data files that are used for genotype calling (.CEL/.IDAT/.CNCHP/other).
2. Text files containing genotype and/or cnv calls, confidence scores, raw and normalized intensities (.CHP/.ALLELE\_SUM/.report.CSV/cn\_segments/.CopyNumber.txt/other).

Please note that one file per sample is the preferred format; however matrixes of intensities, genotypes, and confidence scores for the whole sample set are acceptable as well.

Please see [list of questions](#questions) related to genotype submission, which dbGaP genotyping team may ask before processing the data.

1. **Submitting genotypes and genotype-related data to dbGaP**
2. **Submitting genotype calls/data in individual format**

Genotypes produced using genotyping arrays/marker\_panels usually are submitted in individual format (one file per genotyped sample).

1. **Text format individual genotype data**
2. ***Rows.*** Each individual file contains as many rows as there are markers in the manufacturer’s array annotation/manifest file (for example in Human610-Quadv1\_B.bpm/.csv)
3. ***Columns.*** Example of columns included in a standard CIDR’s genotype report file

(\*-Required columns):

SNP Name (\*)

GC Score (\*)

Allele1 – Forward

Allele2 - Forward

Allele1 – Top (\*)

Allele2 – Top (\*)

Allele1 - Design

Allele2 - Design

Allele1 - AB

Allele2 - AB

Theta

R

X intensity (\*)

Y intensity (\*)

X Raw

Y Raw

B Allele Freq

Log R Ratio

1. **Raw individual genotype data -** dbGaP accepts genotype raw data used to generate genotype calls:
2. .idat files for Illumina arrays (two file per sample)
3. .CEL files for Affymetrix arrays (one file per sample)
4. Other
5. **Genotype meta information**
6. Please provide mapping of each submitted genotype file (raw and text) to the sample level IDs included in 5a\_dbGaP\_SubjectSampleMappingDS file (described in Submission\_Guide\_Instructions.docx). Such FileSampleMapping file should contain at least two columns: (i) FileName and (ii) SampleID
7. MarkerAnnotationFile. Please provide standard annotation/manifest file for the array/platform/marker\_panel used for genotyping.
8. **Submitting CNV data in individual format**

CNV genotypes produced using genotyping arrays usually are submitted in individual format (one file per data type per genotyped sample). Please submit intensities/CNV\_ calls/CNV\_segments and raw/auxiliary data used to generate CNV calls:

1. **Text format individual CNV data**
2. Probe level copy number calls and intensities (.COPYNUM files; one file per sample)

Examples of Indiv\_file\_name..COPYNUM submitted to dbdGap (\* required columns):

SNPID\*

Chromosome\*

Position\*

Log2Ratio\*

AllelicDifference\*

GenotypeCall\*

GenotypeConfidence\*

CNState

GaussianSmooth

LOH

1. Gain/Loss segments (.CNSEGMs; one file per sample)

Examples of Indiv\_file\_name . cn\_segments submitted to dbdGap (\* required columns):

SampleID/individual\_CNCHP\_file\_name

Copy Number State\*

Loss/Gain\*

Chr\*

Cytoband\_Start\_Pos

Cytoband\_End\_Pos

Size(kb)\*

#Markers\*

Avg\_DistBetweenMarkers(kb)

%CNV\_Overlap

Start\_Linear\_Pos\*

End\_Linear\_Position\*

Start\_Marker\*

CNV\_Annotation

1. **Raw individual CNV data -** dbGaP accepts CNV raw data used to generate CNV calls:
2. .CEL files files for Affymetrix arrays (one file per sample)
3. Binary .CNCHP files with probe info for CNV probes (one file per sample)
4. **CNV meta information**
5. Provide FileSampleIDMap (FSM, columns: SAMPLE\_ID; File\_name) mapping each submitted file to the sample ID which is listed in the study’s Subject Sample Mapping File (columns: SUBJECT\_ID;SAMPLE\_ID).

In the FSM, all file names belonging to the same sample are expected to be linked to the same SAMPLE\_ID.

1. Please provide short description of method/algorithm used to generate CNV calls (free formatted text)
2. Please provide any information that may be relevant for QC and data verification
3. **Submitting genotypes in VCF (matrix) format**
4. **Expected specifications for genotypes in VCF format**
5. ***VCF file*** should contain sample ID which are declared in the study **dbGaP Subject Sample Mapping File** (**SSM**) as individual identifiers (the same sample ID should be available in the individual .bam file header).
6. ***FORMAT***. VCF file should be validated using vcftools.
7. ***HEADER***. Header of VCF file should contain:
8. Unambiguously and properly named reference genome sequence used.
9. Unambiguously and properly named dbSNP build (if not dbSNP, then the proper name of the molecular marker database used).
10. Unambiguously and properly named target list/annotation file for projects with targeted sequencing data.
11. Please exclude long internal paths to the individual data files from the header.
12. Indicate if positions 0 or 1-based.
13. ***SNPs/POSITIONS***.
14. RS# should be provided for the positions if available.
15. All .vcf files submitted for the study must be annotated using the same marker annotation file.
16. When applicable please provide target list used for the enrichment as a separate file or url if available.
17. **dbGaP will**
18. Check consistency between sample IDs in VCF and SSM files.
19. Split .vcf by consent groups if needed (using vcftools subset command). Please note that dbGaP team will not merge .vcf files.
20. Check for proper data and header formatting.
21. Check for compatibility with vcftools (using validate and/or conversion to plink).
22. Check for unexpected subject duplicates (by conducting IBD on subset of SNPs) and first degree relations if pedigree data is provided within phenotype dataset.
23. Check for missing call rate per SNP, per sample (may be done by chromosome). SNP filter files will be generated and provided to users (using plink or vcftools).
24. Check minor allele frequency. SNP filter files will be generated and provided to users (using plink or vcftools).
25. Check gender against info provided with phenotype dataset (using plink or vcftools).
26. SNP filter for Mendelian error rate will be generated if sequences from trios will be available (using plink or vcftools).
27. **Submitting imputed genotype data**
28. Imputed genotype data are may be formatted as expected by software used for conducting the imputations (BEAGLE/IMPUTE2/other)
29. Imputed genotype data are expected to be divided by chromosome, by consent group and if needed by population (within each consent)
30. In addition to genotype data, please provide general description of imputed dataset/imputation process.
31. **Submitting genotype QA results**
32. dbGaP accepts any QC metrics/reports produced by PI and/or DCC group(s) on genotype/molecular data.
33. Many dbGAP studies contain GWAS precomputes produced by PI/DCC group(s). Please submit if available.
34. **Submitting to other NCBI databases**
35. **Submitting to dbVar.** If study contains CNV data, in addition to the data released through dbGaP, our team will contact you about submission of CNV regions to the database of Structural Variation (dbVar, <http://www.ncbi.nlm.nih.gov/projects/dbvar>). Most of the required data processing will be done by the dbGaP genotype team on your behalf. Individual CNV regions from each study subject will be disassociated to prevent the possibility of whole genotype reconstruction, which will make possible public release of the data. An investigator may be asked to provide additional study level information required for the standard dbVar submission forms.

1. **Submitting to ProbeDB and to dbSNP.** To facilitate marker information updates the dbGaP team will submit array probes to Probe (<http://www.ncbi.nlm.nih.gov/probe/>) and to dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>). Both databases periodically map submitted probes to a current genome build. After such an update the user will be able to download current genome positions for all array markers at once using array batch ID.

**TIPS**

1. **What to submit**

Preparing data for posting on dbGaP, please consider the “updatability” of the submitted individual level material. Any future updates of submitted genotyped sample sets (addition, removal, or changing consent of the study subjects) will trigger changes in the released genotype components. In most cases, all essential and required genotype information will be updated by the dbGaP team after consulting with you. However, the submitter will be responsible for updating the supplementary individual level data.

Example: In addition to the main genotype dataset, the investigator decided to post a specifically formatted data project with individual genotype calls for the same subject, made with different algorithm/software. Later, after the data were released, some of the subjects changed their consent and/or the investigator came to conclusion that some subjects had to be removed from the study. This would trigger a dbGaP study update. In most cases, updates of the genotype datasets would be conducted by the dbGaP team using your instructions. However, you would be asked to update the supplementary project and re-submit its truncated version.

1. **Questions about submission of genotype data**

List of questions the Investigator will be asked by dbGaP curators regarding the submission of genotypes:

1. Were some subjects genotyped more than once? If yes, were the duplicated samples included in the dbGaP Subject Sample Mapping Data File and if not, where is the file linking Subject IDs to Sample IDs located.
2. Are there “not recommended for GWA” samples and will their genotypes be released? If yes, what is the reason for exclusion and where is the list of these samples located?
3. Do individual genotype files contain within or named using corresponding Sample ID? If not, please indicate which file contains mapping of each individual data file to the corresponding Sample ID.
4. Please provide official name|title for the genotyping array (as indicated in the manufacture’s platform manifest (if commercial array(s)). Otherwise, please provide array manifest with array title, description, and list of markers with marker name, allele 1/allele2 sequence, 5’ and 3’ flank sequences as used on the array (if custom array).
5. Please provide the genome build the SNPs were called on. Please specify forward or reverse strand.

**TERMS**:

**bulk of the study’s genotype data** OR**individual level data files** – refers to the variety of the genotype-containing (report\_CSV/CHP\_ALLELE-SUMMARY/other) and genotype-related (CEL/IDAT/other) data files associated with one sample.

**dbGaP Subject Sample Mapping File** – refers to the file in a standard format described in the dbGaP Submission Guide which contains list of all genotyped samples and corresponding subjects.

As a rule dbGaP Subject Sample Mapping File should include**:**

1. only genotyped subjects to be released;
2. all control samples with corresponding Subject IDs;
3. all duplicated study or control samples (including HapMap) with unique Sample IDs;

dbGaP Subject Sample Mapping File for a study with duplicated samples is supposed to contain a column with unique list of Sample IDs and Column with non-unique list of Subject IDs (samples belonging to the same subject would have the same Subject ID)

dbGaP Subject Sample Mapping File for a study without duplicated samples may contain two identical columns named as SUBJECT\_ID and SAMPLE\_ID.

**dbGaP Subject Consent File** - refers to the file in a standard format described in the dbGaP Submission Guide which contains list of all subjects either “phenotyped”| “genotyped” or both with corresponding consent information. Every study or control (including HapMap) subject mentioned in any study documents must be listed in the dbGaP Subject File.