Submission Guidance Notes

All data submitted to HGVbaseG2P will remain the property of the data generators and/or submitters, and all records will be presented to database with links and acknowledgements leading back to the original data source. Any users who might wish to obtain non-aggregated data will be instructed to make suitable requests to the relevant submitter and their data access authorities.

Submissions can be submitted with embargo dates or conditions attached. We will still immediately process such datasets to ensure the submission is complete and useable, but we will not release the submitted data to the public until instructed to do so.

When submitting genetic association data and/or allele/genotype frequency data to HGVbaseG2P, we require that the utilised Markers are all present in a major public marker/variation database (e.g., dbSNP). If this is not the case, we can assist you in depositing the Markers into a suitable database.

To submit genetic association and/or allele/genotype frequency data, please gather together the information listed below (fields marked * are required). An Excel sheet is provided to help you with this task. Each submission will equate to one Study in HGVbaseG2P, but each Study (i.e., each submission) can include one or more Experiments.

Note: for the future, we are devising a standalone software tool that submitters will be able to download and install locally, which will actively guide them through the process of gathering and checking their data before submitting it. The tool will organize submission content into an XML formatted document that is stored on the submitter's hard disk, gather related information from sites across the internet (e.g., journal citation details, Marker Ids, and Allele specifications), and check for any inconsistencies in the total submission. This will make it simpler for users to assemble and check their submissions with care at their own pace, with the added benefit that they will be able to reuse components (e.g., assay details, clinical materials, and phenotype descriptions) from earlier submissions.

Questions on making submissions should be directed to:
submissions@hgvbaseg2p.org

Study

The Study should be thought of as a databased version of a journal article. Therefore, send in one submission for each set of information you believe tells a single, useful story.

*Study Title	Brief title for the study (free text)
Abstract	Brief summary of the Study and its findings (free text)
Background	Describe the background to the project (free text)
Objectives	Main objective of the study (free text)
Key Results	Important results from the study (free text)
Conclusions	Important conclusions (free text)
Study Design	Specify the Study design (free text)
Acknowledgements	Acknowledgements sentences (free text)
*Author Names	Indexed list of Study authors (1 = corresponding author (CA))
Departments	Indexed list of authors Departments
*Institutions	Indexed list of author institutions (required for CA)
Addresses	Indexed list of authors addresses
*Emails	Indexed list of authors emails (required for CA)
WWW	Web link to the CA's institutional webpage
Pubmed IDs	List of PubMed IDs for any papers describing the Study

Phenotypes

Submissions may include any number of Phenotypes per Study, and these are broken down into the name of the phenotype trait (e.g., 'blood pressure', or 'Alzheimer's Disease') and the method used to test for that phenotype trait (e.g., 'manual use of blood pressure cuff', or 'clinical evaluation of AD using DSM3 criteria'). The observation or result obtained by applying the method is not listed here, but in the Sample Panel and Assayed Panel section.

*Phenotype Name	A name for the Phenotype trait (free text)
*Method Name	A name for the method used to test the Phenotype (free text)
*Method Details	Description of details of the method (free text: include such things as the assayed biological sample, the measured attribute, the circumstance of the test, the time instant and period of operation, the units of measurement, etc)
Description	Additional comments to delineate the Phenotype (free text)

Sample Panels

Submissions may include any number of Sample Panels per Study (but at least one is required), and the details for each of these should be separated one above the other towards the top of the 'Panels' sheet in the Excel submission document.

Sample Panels (continued)

*Sample Panel Name	Unique name for the Sample Panel (free text)
*Description	General description of the Sample Panel (free text)
*Phenotypes Available	List of Phenotypes for which measurements are available for the Sample Panel (list of values from the list of submitted Phenotype Names)
*Phenotype values	List of Phenotype measurements for the Sample Panel (list of categorisations/distributions for the Phenotype Names listed in the last data item: such as 'all taller than 2m', or 'Late-Onset Alzheimer's Disease patients', or 'body weight distribution: mean=110kg, max=150, min=90kg')
*No. of Individuals	Total number of people in the Sample Panel (integer)
No. of Males	Total number of males in the Sample Panel (integer)
No. of Females	Total number of females in the Sample Panel (integer)
Mode of Recruitment	Describe how the individuals were recruited (free text)
Sampling Period	When were the individuals recruited (free text)
Sampling Age Range	How old are the individuals (free text)
Population Info	What 'population' label best fits this Sample Panel (free text)
Geographic Region Info	Where in the world do the individuals live (free text)
Ethnicity Info	What ethnicity are the Sample Panel members (free text)
Environment Info	Describe environment exposures of importance (free text)
DNAs are WGA	Were any DNAs whole genome amplified (all/some/none)

Assayed Panels

Submissions may include any number of Assayed Panels per Study (but at least one case and one control panel are required), and the details for each of these should be separated one above the other on the 'Panels' sheet in the Excel submission document. The complete list of Assayed panel details should be placed below the Sample Panel details.

*Assayed Panel Name	Unique name for the Assayed Panel (free text)
*Description	General description of the Assayed Panel (free text)
*No. of Individuals	Total number of people in the Assayed Panel (integer)
No. of Males	Total number of males in the Assayed Panel (integer)
No. of Females	Total number of females in the Assayed Panel (integer)
*Sample Panel Names	List of Sample Panels from which the Assayed Panel is assembled (one or more Names from the submitted Sample Panel Names)
*Sample Panel No.s	List of number of individuals taken from each of the Sample Panels listed in last data item (one or more integers)
Phenotype Criteria	Phenotype criteria that were used as the basis for choosing which individuals to include in the Assayed Panel (free text)
Environment Criteria	Environment criteria that were used as the basis for choosing which individuals to include in the Assayed Panel (free text)

General info for Genotype Experiments (allele/genotype frequency data sets)

Submissions may include any number of Genotype Experiments per Study, and the general information for each of these should be separated one above the other on a single sheet in the Excel submission document.

*Experiment Title	Brief title for the Genotype Experiment (free text)
*Objective	Summarise the Experiments Objective (free text)
*Outcome	Summarise the Experiments Outcome (free text)
Comments	Further information regarding the Experiment (free text)
*Genotyping Platforms	List of technologies used e.g. AffymetrixXXX
*Individual Data	Is individual level data available upon request? (yes/no)
*Link To Source Data	How to access the source data, with URL if possible (free text)

General info for Analysis Experiments (genetic association data sets)

At present we can only accept data from case-control type association studies, though other approaches (such as TDT) will be supported in the near future. Submissions may include any number of Analysis Experiments per Study, and the general information for each of these should be separated one above the other on a single sheet in the Excel submission document. Analysis Experiments must relate to the genetic basis of one phenotype.

*Experiment Title	Brief title for the Analysis Experiment (free text)
*Phenotype Tested	Phenotype tested for genetic association (one of the list of submitted Phenotype Names)
*Objective	Summarise the Experiments Objective (free text)
*Outcome	Summarise the Experiments Outcome (free text)
Comments	Further information regarding the Experiment (free text)
*Statistics Method	Name of the statistical method(s) employed
Statistics Details	Details of the statistical method(s) employed
*Genotyping Platforms	List of technologies used e.g. AffymetrixXXX
*Individual Data	Is individual level data available upon request (yes/no)
*Link To Source Data	How to access the source data, with URL if possible (free text)

Frequency and Association Data

Once all the above information has been entered into the appropriate sheets on the Excel submission document, moved to the Experiment Data sheet and paste in one large table of aggregated results (i.e., no individual level results) comprising all the allele/genotype frequency data (for all types of Experiments) plus the corresponding phenotype association p-values (if you're submitting Analysis Experiments).

Frequency and Association Data (continued)

*Experiment Title — The Experiment for that data row (one of the submitted Experiment Titles)

*Experiment Type — Type of Experiment in that data row ('Genotype Experiment', or 'Analysis Experiment')

MARKER INFORMATION:

*dbSNP rsID — Universally recognised Marker ID (dbSNP rsID, or alternative)

*10bp 5' of Alleles — 'upstream' flanking DNA sequence, 10 bases long, from the same DNA strand as the alleles (based upon IUPAC codes)

*Allele 1 — DNA sequence of the alleles you wish to refer to as '1' (based upon IUPAC codes or the underscore character '_' for deletion alleles)

*Allele 2 — DNA sequence of the alleles you wish to refer to as '2' (based upon IUPAC codes or the underscore character '_' for deletion alleles)

*10bp 3' of Alleles — 'downstream' flanking DNA sequence, 10 bases long, from the same DNA strand as the alleles (based upon IUPAC codes)

CONTROL RESULTS (required for all types of Experiments):

*Assayed Panel Name — Assayed Panel used as the control materials (one of the names from the submitted Assayed Panel Names)

*Genotype_11 count — Number of individuals amongst the controls found to carry the '11' genotype, or the '1' genotype for X-chromosome markers in males (integer)

*Genotype_12 count — Number of individuals amongst the controls found to carry the '12' genotype (integer)

*Genotype_22 count — Number of individuals amongst the controls found to carry the '22' genotype, or the '2' genotype for X-chromosome markers in males (integer)

Allele_1 count — Number of chromosomes amongst the controls found to carry Allele_1 (integer)

Allele_2 count — Number of chromosomes amongst the controls found to carry Allele_2 (integer)

CASE RESULTS (only required for Analysis Experiments):

*Assayed Panel Name — Assayed Panel used as the case materials (one of the names from the submitted Assayed Panel Names)

*Genotype_11 count — Number of individuals amongst the cases found to carry the '11' genotype, or the '1' genotype for X-chromosome markers in males (integer)

*Genotype_12 count — Number of individuals amongst the cases found to carry the '12' genotype (integer)

*Genotype_22 count — Number of individuals amongst the cases found to carry the '22' genotype, or the '2' genotype for X-chromosome markers in males (integer)

Allele_1 count — Number of chromosomes amongst the cases found to carry Allele_1 (integer)

Allele_2 count — Number of chromosomes amongst the cases found to carry Allele_2 (integer)

ASSOCIATION P-VALUES (only required for Analysis Experiments):

*Allele Association — The uncorrected case-control p-value statistic for an analysis of allele association (number expressed in log-10)

*Genotype Association — The uncorrected case-control p-value statistic for an analysis of genotype association (number expressed in log-10)