HEALTH-F4-2007-200754

http://ww.gen2phen.org

D7.1 dbSNP-lite Established

WP7 – Data Flows

V0.4
Final

Lead beneficiary: UNIMAN
Date: 08/03/2010
Nature: Other
Dissemination level: PU (Public)

TABLE OF CONTENTS

Document Information

Grant Agreement Number	HEALTH-F4-2007-200754	**Acronym**	GEN2PHEN
Full title	Genotype-To-Phenotype Databases: A Holistic Solution		
Project URL	http://www.gen2phen.org		
EU Project officer	Frederick Marcus (Frederick.Marcus@ec.europa.eu)		

Deliverable	**Number**	7.1	**Title**	dbSNP-lite Established
Work package	**Number**	7	**Title**	Data Flows

Delivery date	**Contractual**	Month 24	**Actual**	15/02/2010
Status	Version 0.4		final ☑	
Nature	Report ☐ Prototype☐ Other ☑			
Dissemination Level	Public ☑ Confidential☐			

Authors (Partner)	Gudmundur. A. Thorisson (ULEIC)			
Responsible Author	Gudmundur A. Thorisson		**Email**	gt50@le.ac.uk
	Partner	ULEIC	**Phone**	+44 116 239 7723

Document History

Name	Date	Version	Description
P. Sarmah (CSIR)	01/01/2010	0.1	Initial draft
G. A. Thorisson (ULEIC)	15/01/2010	0.2	Draft
A. Devereau (UNIMAN)	01/02/2010	0.3	Internal review
G. A. Thorisson (ULEIC)	01/02/2010	0.4	Amended draft

Definitions

- Partners of the GEN2PHEN Consortium are referred to herein according to the following codes:

  ULEIC – University of Leicester (UK) – Coordinator
  EMBL – European Molecular Biology Laboratory (Germany) – Beneficiary
  FIMIM – Fundació IMIM (Spain) – Beneficiary
  LUMC – Leiden University Medical Center (Netherlands) – Beneficiary
  INSERM – Institut National de la Santé et de la Recherche Médicale (France) – Beneficiary
  KI – Karolinska Institutet (Sweden) – Beneficiary
  FORTH – Foundation for Research and Tecnology Hellas (Greece) – Beneficiary
  CEA – Comissariat à l'Energie Atomique (France) – Beneficiary
  EMC – Erasmus Universitair Medisch Centrum Rotterdam (Netherlands) – Beneficiary
  UH.FGC – Helsingin Yliopisto (Finland) – Beneficiary
  UAVR – Universidade de Aveiro (Portugal) – Beneficiary
  UWC – University of the Western Cape (South Africa) – Beneficiary
  CSIR – Council of Scientific and Industrial Research (India) – Beneficiary
  SIB – Swiss Institute of Bioinformatics (Switzerland) – Beneficiary
  UNIMAN – The University of Manchester (UK) – Beneficiary
  BIOBASE – BioBase GmbH. (Germany) – Beneficiary
  deCODE – Islensk Erfoagreining EH (Iceland) – Beneficiary
  PHENO – Phenosystems S.A. (Belgium) – Beneficiary
  BCP – Biocomputing Platforms Ltd. Oy (Finland) – Beneficiary

- **Grant Agreement**: The agreement signed between the beneficiaries and the European Commission for the undertaking of the GEN2PHEN project (HEALTH-200754).
- **Project**: The sum of all activities carried out in the framework of the Grant Agreement by the Consortium.
- **Work plan**: Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out for the GEN2PHEN project, as specified in Annex I to the Grant Agreement.
- **Consortium**: The GEN2PHEN Consortium, conformed by the above-mentioned legal entities.
- **Consortium agreement**: agreement concluded amongst GEN2PHEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

1. INTRODUCTION

Recent years have seen huge advances in genetic epidemiology studies into complex, common diseases, largely enabled by the emergence of high-throughput, microarray-based genotyping platforms. These recently-developed technologies facilitate simultaneous interrogation of hundreds of thousands of single nucleotide polymorphisms (SNPs) in the genome at very low cost per sample, and have thus enabled large-scale, genome-wide association studies (GWAS) involving up to tens of thousands of participants[1].
The rapid evolution of SNP genotyping platforms in the past decade is founded on exponential increases in available information regarding the extent of SNPs, insertions/deletions (indels) and other small sequence variants in the human genome. Large-scale discovery projects[2,3] have created vast genome-wide collections of these variants,a foundation upon which the Haplotype Map Project (HapMap: http://www.hapmap.org)[4] has been built to create a map of the patterns of common SNP variation across the genome. The detailed information on the extent and patterns of genetic variation resulting from these large-scale discovery projects has enabled the design of SNP microarrays which directly or indirectly capture the majority of known common genetic variation.

1.1. dbSNP: a reference catalog of simple sequence variation

A key consideration in the analysis and interpretation of GWAS findings is existing knowledge of genetic variation.The database of Single Nucleotide Polymorphisms (dbSNP: http://www.ncbi.nlm.nih.gov/SNP/)[5], developed and operated by the US National Center for Biotechnology Information (NCBI) since the late 1990s, is the *de facto* reference source of information on simple, sequence-level variation in human and several other organisms. Many G2P data resources which are focused on association studies rely on dbSNP data as a reference to contextualize study findings. Some of these integrate their data with dbSNP via simple Web hyperlinks from a report page to the dbSNP website. Others utilize the NCBI Entrez Programmatic Utilities (eutils: http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) to retrieve dbSNP variation data via web services as needed, for analysis or display on their websites. Still others require a local copy of a core subset, or even a complete mirror, of dbSNP contents in order to support local computationally-intensive analyses and other processing.

1.2. A changing picture of genetic variation

As with any other reference sources of biological data, each of these modes for utilising dbSNP data requires dealing with changes in the source database, in particular in the case of locally stored copies of the data. The reference information layer of known genetic variation as archived by dbSNP is dynamic and changes constantly over time as new data are gathered and existing records are altered. Ongoing variation discovery projects such as the 1000 Genomes Project (http://www.1000genomes.org) identify new and confirm suspected variant sites in the genome, and thus extend and enrich the dbSNP catalog. The reference genome sequence improves in

quality, so genome mapping information in dbSNP changes with each new genome build. Erroneous dbSNP submissions are corrected or deleted.

As a result of these and other changes, association study findings can be expected to gradually become inconsistent with data in the reference dbSNP archive. For example, some types of changes may invalidate earlier assumptions regarding genotyping assay designs or allele calling algorithms in generation of primary GWAS genotype data. Other changes may affect downstream analysis in subtle ways, for example changes in reported allele strand orientation. This can complicate comparison and integration of association study findings from different points in time, generated in the context of different releases of dbSNP. Maintaining a comprehensive view of these changes in the primary dbSNP variation archive is therefore of key importance in Work Package 7 'Data Flows', which is concerned with the flow of data into and between the databases and tools that form the data and analysis structure being developed by the GEN2PHEN project.

1.3. Motivation for dbSNP-lite

For the reasons explained, tracking changes in dbSNP content is an important activity for the Human Genome Variation database of Genotype and Phenotype (HGVbaseG2P:http://www.hgvbaseg2p.org)[6], which is being developed as part of WP5. A chief concern in HGVbaseG2P operations is to maintain a consistent link between association study data and a basal data layer of genetic marker data, such that references to primary marker information in the former can be properly updated to match changes in the latter. However, the marker provenance information provided by dbSNP is not adequate for the detailed level of marker revision tracking deemed necessary for this task. In this report we describe a key outcome of WP activity 7.2 'Populating Genomics Databases': the design and implementation of deliverable D7.1 'dbSNP-lite' data processing pipeline for extracting a core set of data from dbSNP and synchronising with a local copy of marker data in HGVbaseG2P extracted from a previous dbSNP release.A key function of this tool is to compare marker entries in a new dbSNP release with existing marker entries in HGVbaseG2P and highlight a key set of changes(e.g. SNP mergers, splits, corrections, deletions) relative to the previous dbSNP release.
The output of the dbSNP-lite tool is a "slim" version - or abstraction of core elements – from a complete dbSNP release in a standard format, enhanced with revision information for each marker describing the changes, if any, that were found. These data may have broader utility for others who wish to utilise dbSNP data in a similar way to HGVbaseG2P, and in recognition of this the tool is designed to be extended to extract additional data elements from dbSNP if required. Furthermore, capabilities for handling other reference variation sources, such as the Database of Genomic Variants (DGV: http://projects.tcag.ca/variation/) and dbVar (http://www.ncbi.nlm.nih.gov/projects/dbvar/) will be added in the near future.

2. Methodology

2.1. Overall design and architecture

A two step approach, illustrated in Figure 1, was chosen for synchronising dbSNP with HGVbaseG2P. In the first step, the dbSNP-lite application processes dbSNP bulk datafiles and generates a set of standard GFF3-formatted output files. In the second step, the GFF3 files are loaded into the HGVbaseG2P Marker database. The rationale for this strategy was as follows. First, each of the two main software component could be focused on one task, and thus could be made simpler and easier to maintain (compared to a monolithic, more complex application). Second, a separate data loading step carries less risk of partial or incorrect data being imported into the database (since the full set of intermediary files from the processing stage can be checked beforehand). Third, the intermediary files may be useful to others by themselves, and there was a desire to make the dbSNP-lite tool and its output more broadly useful, even without the presence of the full HGVbaseG2P system.

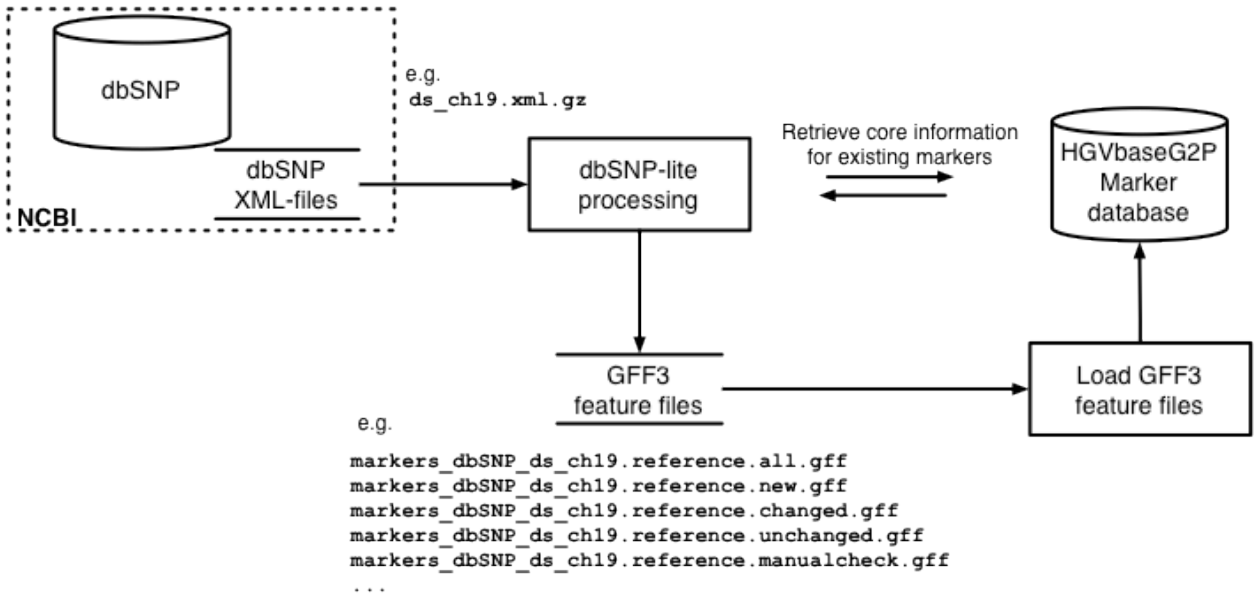


Figure 1: Summary view of the dbSNP-lite workflow for synchronising dbSNP with HGVbaseG2P

2.2. Implementing the dbSNP-lite application

The dbSNP-lite tool was created as a joint project undertaken by representatives of GEN2PHEN partners CSIR (P. Sarmah) and ULEIC (G.A. Thorisson). The application is implemented in the Perl programming language and comprises a number of command-line scripts and several library modules. Common processing logic resides in a master superclass whilst datasource-specific functionality is implemented in derived classes. At present only a derived class for dbSNP has

been implemented, but modules for other sources of marker data can be easily added. Depending on the mode of operation, several helper classes from the main HGVbaseG2P codebase are also utilised for database querying. However, these are not required for the standalone mode of the application.

2.3. Input data for analysis

Each release of the dbSNP database is made available in bulk in several formats, including flat files, a variety of gene- and chromosome-oriented reports, full database table dumps and as a set of XML-files ordered by chromosome. For our purposes, the XML-dumps were considered sufficiently comprehensive,and moreconsistent and convenient to process than the other bulk download options.To prepare for dbSNP-lite processing, all available XMLs for the current dbSNP build (b130) were downloaded from the NCBI FTP-site (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/XML/) onto the HGVbaseG2P Unix server.

dbSNP organisation and build procedure. The central unit of organisation in dbSNP is the reference SNP cluster, usually referred to as a refSNP or rs#. An rs# represents a site in the genome, defined by a pair of 5' upstream and 3' downstream flanking sequences, which has been shown to vary between individuals. The stable, accessioned rs# record is created from one or more submitter SNP entries, or ss#'s, submitted by variation discovery projects large and small, which are clustered together based on sequence similarity and common mapping to a genomic contig. The final, reference list of alleles for the reference SNP entry is the combined set of alleles reported in each ss#.The number and source of the constituent ss#'s in a given rs#, as well as population data and other information, are used to assign validation status to the rs#, indicating the likelihood of the polymorphism being a *bona fide*variation and not an experimental artifact.

dbSNP reclustering. The NCBI prepares a new dbSNP build approx. 1-2 times per year by re-clustering all available ss#'s to create a new reference set of non-redundant rs#'s. During this build procedure, new ss#'ssubmitted to dbSNP since the last release are either clustered together with existing rs#'s, or else are used to seed new rs# clusters at sites in the genome not previously known to vary. Occasionally this results in changes at the rs# level (see below), but unlike GenBank and many other primary archives dbSNP does not employ a versioning and archiving for tracking such changes and making previous versions of changed rs#'s accessible.

rs# mergers and deletions. A side-product of evolution and improvements of the dbSNP re-clustering procedure and changing genome assemblies is that sometimes two or more rs# are found to represent the same polymorphic site in the genome. This results in a so-called refSNP 'merge' event between the co-located rs#'s, whereby the rs# with the higher number is merged into the rs# with the lower number(e.g. rs58061040 => rs626358) and subsequently deleted. rs#'s are also removed from the database if the underlying submitter-provided ss# entries are withdrawn for some reason. As with changes at the rs# level, deleted markers in dbSNP are

likewise not archived and kept accessible; instead they are removed completely from database releases (but see Discussion for exceptions to this).

2.4. Running dbSNP-lite

The dbSNP-lite application can be run in one of two modes. The first mode is intended for non-HGVbaseG2P parties who do not have access to the main HGVbaseG2P database and software suite. The tool can be run in this 'naive' mode as follows on the command-line, for each input XML-file:

```
markerGFFFromAny.pl -naive -s dbSNP -l b130 [XML-file]
```

Further information on dbSNP-lite operation can be found in the command-line documentation, accessible (as per Unix convention) by running the tool with the `--help` option.

2.5. Core processing of dbSNP entries

For each reference SNP entry foundin the input XML-file (the <Rs> element), a core subset of available marker information is extracted:

- dbSNP rs# accession
- Variation class (e.g. "SNP", "indel")
- Validation code (e.g. "byFrequency", "byHapMap")
- List of reported alleles (e.g. "A/C", "-/TTG")
- 5' upstream and 3' downstream flanking sequences (30bp on either side)
- Cross-references to other NCBI resources, such as dbSTS and LocusLink
- Mapping information for all available genome sequence assemblies

The second, full mode for running the tool requires a configuration file to be provided which specifies connection parameters for the target database:

```
markerGFFFromAny.pl -c conf/hgvbase.conf -s dbSNP -l b130 [XML-file]
```

At present, this mode requires direct access to the HGVbaseG2P database which is not available to external users at this time, as well as a custom database API which is not provided with this first version of dbSNP-lite (see also Discussion).

2.6. Detecting and validating changes in core marker information

In addition to the data extraction step, for each rs# entry the workflow depicted in Figure 2 is executed. The target HGVbaseG2P Marker database is queried using the dbSNP rs# accession for

the marker. If the marker is not found in HGVbaseG2P, it as assumed to be a new marker. Otherwise the information from the XML-file is compared with information from the database and a series of checks are undertaken in order to identify which of a specific set of changes, if any, have occurred since the last dbSNP release was incorporated into HGVbaseG2P. Once this procedure has been completed for all markers in the input XML-files, each marker has been placed into one of the following categories:

New: Marker is not present in HGVbaseG2P.
Unchanged: Marker is present in HGVbaseG2P and no changes were detected.
Changed: Marker is present in HGVbaseG2P. One or more changes have occurred since the last source database build was processed and these could be reconciled automatically.
ManualCheck: Marker is present in HGVbaseG2P. One or more changes have occurred since the last source database build and these could NOT be reconciled automatically.

A diagram depicting the workflow followed to identify changes is provided in Annex I. The specific changes are further described below:

Change in variation class:a simple string comparison isused to detect changes in dbSNP-assigned variation class, such as a change from ‘snp’ to ‘mixed’ if an ss# reporting an insertion allele has been added to a rs# cluster where previously only single-nucleotide alleles had been reported.

Change in flanking sequences: occasionally the flanking sequences for established rs# entriesis altered. This is typically due to changes in the set of ss#’s underlying the cluster, such as when new ss#’s have been clustered to the rs#. The ss# with the longest flanks is used as the cluster exemplar, and the master rs# flanks are changed to match the exemplar ss# if needed. Depending on the strand orientation of the exemplar ss# relative to the rs#, this process has in the past sometimes resulted in reverse-complementation of flanking sequences, or ‘strand flip’, for the rs# cluster. dbSNP-lite checks for flank changes and strand flips by aligning rs# flanking sequences in the current dbSNP build with those from the previous build.

Change in reported alleles: as new ss#’s are clustered into existing rs# clusters, previously known alleles are confirmed with independent observations or new alleles are added to the rs#. dbSNP-lite identifies these changes and checks that they are valid, taking into account potential reverse-complementation of the flanking sequences.

2.7. dbSNP-lite output

The output of dbSNP-lite is a set of standard GFF3-formatted feature files which contain the "lite" representation of dbSNP. A feature file labelled "all" is created for each input file per sequence assembly, containing all marker information and mappings but no change information. This file – the only output produced if ‘naive’ mode was used - is provided as a convenience for users who only require a lite version of the current dbSNPand are not concerned with changes since the last build.

If the second, non-naive mode was used, a further set of feature files are created for each change category listed above. For example, new markers which are mapped to Chr18 in the reference sequence assembly are printed to the output file **`markers_dbSNP_ds_ch18.reference.new.gff`**, and existing markers with no changes mapping to Chr1 in the Celera assembly are printed to **`markers_dbSNP_ds_ch1.Celera.unchanged.gff`**.
As per the standard GFF3 specification(http://www.sequenceontology.org/gff3.shtml), the chromosome, genomic coordinates, source, type and strand are placed in the appropriate columns in the tab-delimited feature file. The variation type is specified as a standard Sequence Ontology (SO) terms, which are mapped to the non-standard dbSNP classification as follows:

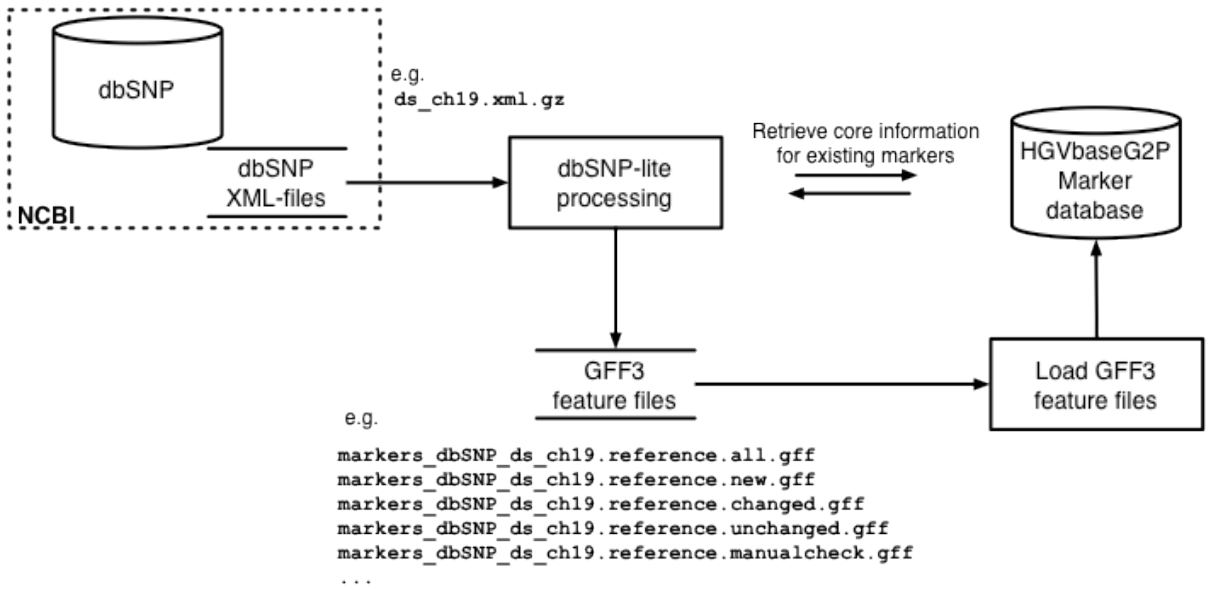


Figure 2: Core dbSNP-lite processing workflow

dbSNP variation class	SO term
snp	SNP
in-del	indel
heterozygous	complex_substitution
microsatellite	tandem_repeat
named-locus	complex_substitution
no-variation	novar
mixed	complex_substitution
CopyNumber	CNV
Inversion	chromosomal_inversion
InversionBreakpoint	inversion_breakpoint

Other marker information is encoded as attribute-value pairs in the 9th column, as per the specification. If a marker has multiple mappings within an assembly or to alternative assemblies, this core marker information is duplicated across the respective chromosome feature files as required. An example illustrating the organisation of marker data in a GFF3-formatted feature file is provided in Annex II.

2.8. Loading marker data into the target database

After running the dbSNP-lite tool to produce the GFF3-formatted feature files, the feature files can be loaded into the target marker database or utilised in other ways. A variety of software tools are available for processing GFF3 files, such as those provided in the BioPerl toolkit (http://www.bioperl.org)[8]. By reusing and extending one of these tools - the `Bio::DB::SeqFeature::Store` feature database and software library - we were able to quickly create a GFF3-loader tool for processing the marker feature files to meet our ends: namely importing both the core marker information and standard feature data into the HGVbaseG2P Marker database. This tool will be released as part of the HGVbaseG2P software suite at later date.

2.9. Checking for deleted markers in source database

As noted above, rs# entries do not appear in bulk dbSNP downloads after having been removed or merged with other rs#'s. This lack of archiving of deleted markers means that a separate procedure is needed to identify which rs#'s now present in HGVbaseG2P are no longer in dbSNP for either of these two reasons. This procedure requires a connection to the HGVbaseG2P Marker database and additional software:

Merged markers. dbSNP provides a running log of merge events in their database table dump. The file `RsMergeArch.bcp.gz` containing this information was downloaded from the dbSNP FTP-site and processed with a custom script in the HGVbaseG2P toolkit. This adds to the HGVbaseG2P Marker database a revision history entry for every rs# which has been merged with another rs# (and subsequently deleted in dbSNP), and creates link between the two entries, effectively duplicating the dbSNP merge history in HGVbaseG2P.

Deleted markers. Following the loading of marker feature files into the HGVbaseG2P Marker database, a final 'legacy' check is undertaken. This is a simple procedure which retrieves all markers which, according to a timestamp-flag, did not feature in the current dbSNP release. For each of these markers, if the rs# is not already logged as deleted due to a merge event, it is flagged as 'dead'.

Importantly, in either of the two scenarios above the affected rs# is **not** deleted from HGVbaseG2P but archived.The net effect of this is that an archive of deleted rs# entries is maintained in HGVbaseG2P, even if they no longer appear in dbSNP itself. This information can

be used for rs# lookups or validation, for example when association study data loaded into HGVbaseG2P refers to rs# identifiers for deleted markers.

3. Results

3.1. Processing dbSNP b130 against b129

The full XML-release of dbSNP 130 was processed against the contents of the HGVbaseG2P Marker database containing data from dbSNP b129. A summary of the results is shown in Figure 3.The majority of the total 17.8M markers in dbSNP remain unchanged (~80%), with a sizeable influx of new markers (~20%) and only ~0.2% markers displaying changes between releases. 5,885 markers were found to have inconsistencies in flanking sequences or allele assignments, and were provisionally excluded from further processing.

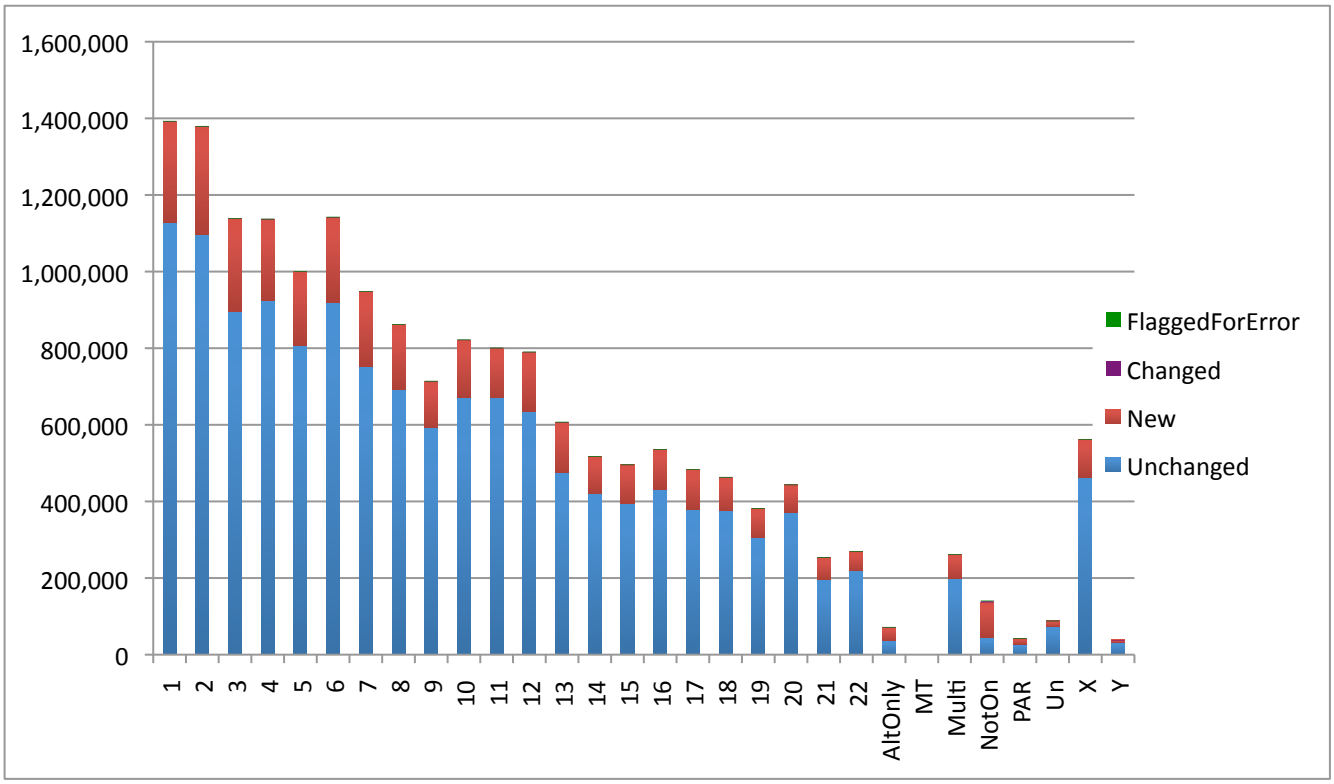


Figure 3 Summary of changes in dbSNP b130 compared to b129 data available in dbSNP (see Annex III for underlying data). Y-axis shows no. rs# entries. X-axis is sorted by dbSNP XML-file labels.AltOnly=rs#'s mapped to a non-reference, alternative genome assembly only; Multi=rs#'s mapped to multiple chromosomes in the reference assembly; MT=rs#'s mapped to the mitochondrion; NotOn=rs#'s not mapped to any chromosome in any assembly; Un=mapped to a chromosome in the reference assembly but no definite sequence coordinates; PAR=rs#'s mapped to the pseudo-autosomal region on ChrY.

3.2. Deleted and merged dbSNP markers in HGVbaseG2P

All feature files, except those labelled 'manualcheck', were loaded into the HGVbaseG2P Marker database using the HGVbaseG2P GFF3-loader utility. dbSNP merge history information was subsequently added to the database, followed by the 'legacy' check as described. The results from this processing are still being analysed and validated, and could therefore not included in this report.

4. Discussion

4.1. New and changed content in dbSNP b130

Since the previous dbSNP b129 release, ~3.5M new rs# entries have been created from data submitted by variation discovery projects, including the 1000 Genomes Project (http://www.1000genomes.org). All these new markers have now been added to the HGVbaseG2P Marker database. Out of the ~14.2M markers already present in HGVbaseG2P, only ~35,000, or ~0.2%, exhibited changes compared to the previous b129. This is a substantial reduction from previous analyses comparing b129 with the previous b128, where ~1% of rs#'s were found to have changed between releases (data not shown). A total of 5,885 rs#'s were found to have changes which could not be reconciled in automated fashion. A full analysis of all ~35,000 markers with changes could not be completed in time for inclusion in this report. Preliminary analysis based on cross-checking rs# lists with the HGVbaseG2P Study database indicates that 10-20 GWAS-tested markers may be affected.

4.2. Limitations of dbSNP-lite and suggestions for future work

In its present form, the dbSNP-lite software implemention has some limitations. This first version of the software is capable of processing only dbSNP-data, but ideally other reference sources of variation data should be supported as well. Some main sources of interest include DGV and the new dbVar database for structural variation (http://www.ncbi.nlm.nih.gov/projects/dbvar/) currently being constructed by the NCBI. Genome mapping information is presently handled in a simplified way, with existing mapping data in HGVbaseG2P simply replaced with mapping data from the new dbSNP release. Here a more intelligent procedure for verifying that marker mappings to the same genome assembly remain unchanged between dbSNP builds may be useful. The software implementation, albeit already modular and extensible, could be enhanced with a simple "plugin" architecture which would make it much easier to extend the package, e.g. for extracting additional fields from the source datafiles.

4.3. Source code and data availability

The sourcecode and documentation for dbSNP-lite are provided on the accompanying CDs. As explained above, this initial version of the tool is limited to running in 'naive' mode which

generates the “lite” version of dbSNP with no change information. A full version of the tool will be released later in the year, as part of the full HGVbaseG2P sourcecode distribution.

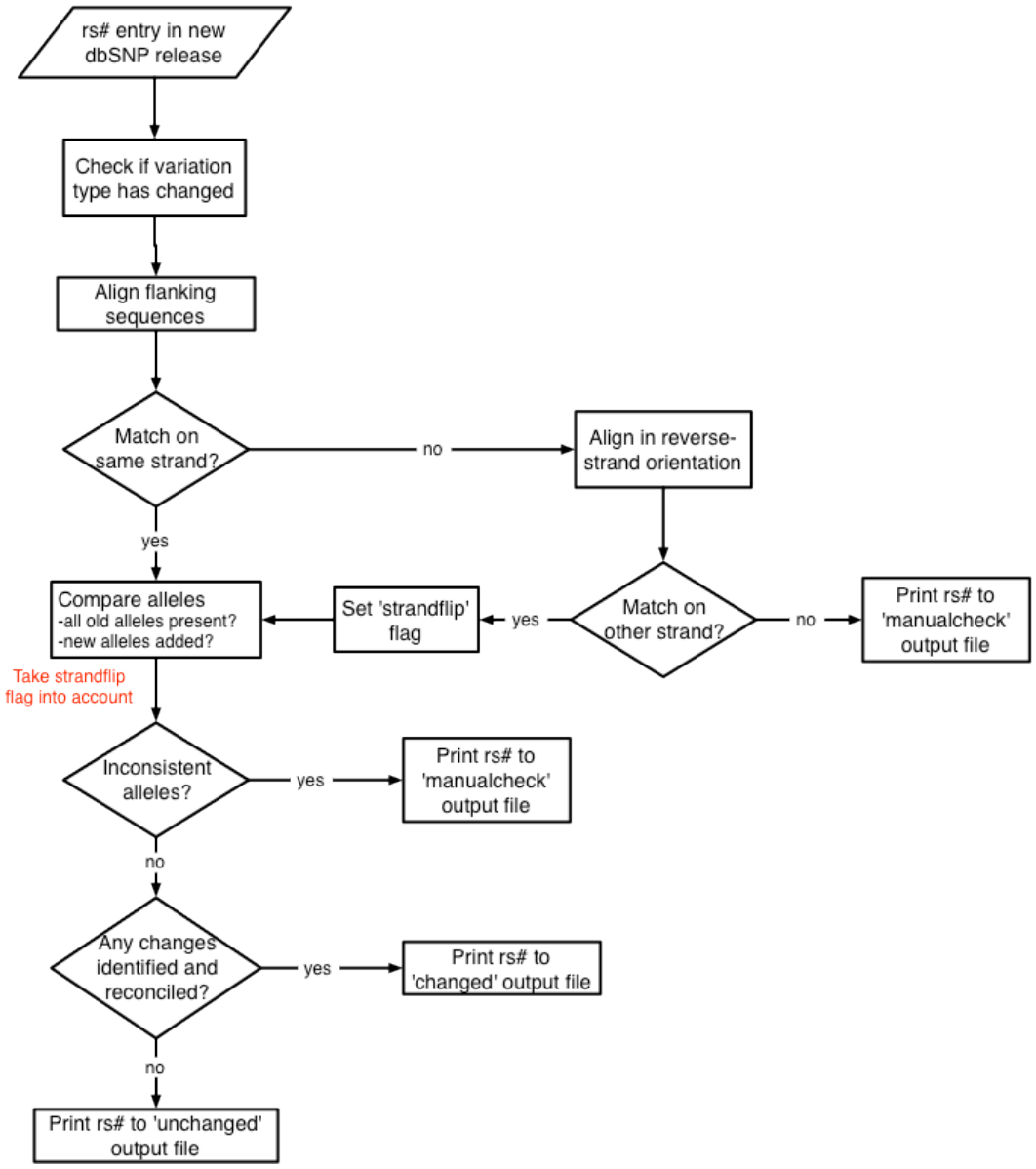
GFF3 feature files for all genome assemblies, as produced by running this version of the tool, is also provided on the CDs. The latest version of these datafiles and the full, enhanced GFF3 files (with change information) will be available shortly at this location on the HGVbaseG2P website: http://www.hgvbaseg2p.org/download/markers/

References

1. Kim, S. & Misra, A. SNP Genotyping: Technologies and Biomedical Applications. *Annu. Rev. Biomed. Eng.* **9**, 289-320 (2007). doi:10.1146/annurev.bioeng.9.060906.152037
2. The International SNP Map Working Group A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928—933 (2001). doi:10.1038/35057149
3. Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–6 (2000).
4. The International HapMap Consortium A haplotype map of the human genome. *Nature* **437**, 1299–320 (2005). doi:10.1038/nature04226
5. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308—11 (2001).
6. Thorisson, G. et al. HGVbaseG2P: a central genetic association database. *Nucleic Acids Res* **37**, D797–802 (2008). doi:10.1093/nar/gkn748
7. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133—141 (2008). doi:10.1016/j.tig.2007.12.007
8. Stajich, J.E. et al. The Bioperl Toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611—8 (2002).

ANNEXES

Annex I – Supplementary workflow diagrams



HGVbaseG2P new dbSNP build comparison/validation workflow

Annex II – Example marker feature data in GFF3 format

[Note that the long GFF3-lines, one per chromosome, are wrapped to multiple lines to fit into this document:

```
##gff-version 3
##genome-build NCBI B36
##species
http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?name=Homo+sapiens
# Timestamp=2010-01-13 10:50:57
# Label=none
# MarkerCategory=new
# AssemblyName=reference
Chr10 dbSNP SNP   114547453   114547453   .     -     .
      ID=rs62657161;Name=rs62657161;upstream30bp=AGCCAATCTGTGATGACGTGAGGATTTTT
T;downstream30bp=TTTGTTTGTTTTCTCACCTTTTTAGTGCCT;alleles=(G):(T);mapweight=uniq
ue-in-contig
Chr10 dbSNP indel 13963268    13963268    .     -     .
      ID=rs62658160;Name=rs62658160;upstream30bp=TCTAGTGAGACCCCCATCTCTAAAAAAAA
A;downstream30bp=AAAAAAAAAAATTAGCTGGGCATGGTGGTG;alleles=(A):(_);mapweight=uniq
ue-in-contig
Chr10 dbSNP indel 12832688    12832688    .     -     .
      ID=rs62658963;Name=rs62658963;upstream30bp=AGGGGATGGGGAAAGTCAAGAATGAATGA
A;downstream30bp=TGAATGAATGATGAATGAATGAATGAATGA;alleles=(TGAA):(_);mapweight=u
nique-in-contig
Chr10 dbSNP indel 124902097   124902097   .     -     .
      ID=rs62659661;Name=rs62659661;upstream30bp=TTCTCACCTCTGTGCCAGTCTTTTGAAGA
A;downstream30bp=ACACACATGGTTCCTCTGACTTAAAGGCTT;alleles=(AC):(_);mapweight=uni
que-in-contig
```

Annex III – Summary of changes in dbSNP b130 vs b129 in HGVbaseG2P

Chromosome	Unchanged	New	Changed	FlaggedForError	Total
1	1,127,418	264,067	1,646	287	1,393,418
2	1,096,022	282,148	1,792	204	1,380,166
3	896,737	240,595	1,543	239	1,139,114
4	924,960	210,716	2,141	120	1,137,937
5	807,799	191,746	1,770	183	1,001,498
6	920,476	220,498	1,460	813	1,143,247
7	751,543	195,878	1,299	137	948,857
8	692,306	168,800	1,442	93	862,641
9	592,009	121,581	1,011	103	714,704
10	671,171	150,151	1,163	178	822,663
11	671,445	128,654	1,075	134	801,308
12	635,917	153,812	1,069	117	790,915
13	475,310	131,302	841	566	608,019
14	420,251	97,329	870	58	518,508
15	395,041	100,636	681	70	496,428
16	432,096	103,288	780	68	536,232
17	380,174	103,051	715	132	484,072
18	375,274	87,759	816	62	463,911
19	307,025	74,973	523	47	382,568
20	371,729	71,473	855	34	444,091
21	195,622	58,803	400	30	254,855
22	218,680	50,817	405	23	269,925
AltOnly	36,692	35,381	33	344	72,450
MT	659	11	0	0	670
Multi	198,856	63,297	82	121	262,356
NotOn	44,907	91,402	3,409	1,652	141,370
PAR	27,678	14,995	39	13	42,725
Un	73,470	15,274	6	8	88,758
X	461,885	99,725	629	49	562,288
Y	32,875	5,456	9	0	38,340
Total	14,236,027	3,533,618	28,504	5,885	17,804,034