Stade et al. BMC Genomics 2014, 15(Suppl 4):S8
http://www.biomedcentral.com/1471-2164/15/S4/S8

BMC Genomics

RESEARCH Open Access

GrabBlur - a framework to facilitate the secure exchange of whole-exome and -genome SNV data using VCF files

Björn Stade[1*], Dominik Seelow[2], Ingo Thomsen[1], Michael Krawczak[3,4], Andre Franke[1*]

From SNP-SIG 2013: Identification and annotation of genetic variants in the context of structure, function, and disease
Berlin, Germany. 19 July 2013

Abstract

Background: Next Generation Sequencing (NGS) of whole exomes or genomes is increasingly being used in human genetic research and diagnostics. Sharing NGS data with third parties can help physicians and researchers to identify causative or predisposing mutations for a specific sample of interest more efficiently. In many cases, however, the exchange of such data may collide with data privacy regulations. GrabBlur is a newly developed tool to aggregate and share NGS-derived single nucleotide variant (SNV) data in a public database, keeping individual samples unidentifiable. In contrast to other currently existing SNV databases, GrabBlur includes phenotypic information and contact details of the submitter of a given database entry. By means of GrabBlur human geneticists can securely and easily share SNV data from resequencing projects. GrabBlur can ease the interpretation of SNV data by offering basic annotations, genotype frequencies and in particular phenotypic information - given that this information was shared - for the SNV of interest.

Tool description: GrabBlur facilitates the combination of phenotypic and NGS data (VCF files) via a local interface or command line operations. Data submissions may include HPO (Human Phenotype Ontology) terms, other trait descriptions, NGS technology information and the identity of the submitter. Most of this information is optional and its provision at the discretion of the submitter. Upon initial intake, GrabBlur merges and aggregates all sample-specific data. If a certain SNV is rare, the sample-specific information is replaced with the submitter identity. Generally, all data in GrabBlur are highly aggregated so that they can be shared with others while ensuring maximum privacy. Thus, it is impossible to reconstruct complete exomes or genomes from the database or to re-identify single individuals. After the individual information has been sufficiently "blurred", the data can be uploaded into a publicly accessible domain where aggregated genotypes are provided alongside phenotypic information. A web interface allows querying the database and the extraction of gene-wise SNV information. If an interesting SNV is found, the interrogator can get in contact with the submitter to exchange further information on the carrier and clarify, for example, whether the latter's phenotype matches with phenotype of their own patient.

Background

Since the introduction in 2005, Next Generation DNA Sequencing (NGS) has been used successfully in numerous research projects [1]. Meanwhile, further technological advances have reduced the per base pair sequencing costs dramatically, thereby allowing more and more molecular diagnostics laboratories to screen the complete exome of individual patients with an apparently inherited disease for causative mutations [2]. Indeed, exome sequencing has already started to revolutionize diagnostic genetic testing [3][4]. However, pertinent data privacy law, the type of informed consent declarations used and limited genetic

* Correspondence: b.stade@ikmb.uni-kiel.de; a.franke@mucosa.de
[1]Institute of Clinical Molecular Biology, Christian-Albrechts University of Kiel, Schittenhelmstraße 12, 24105 Kiel, Germany
Full list of author information is available at the end of the article

BioMed Central

© 2014 Stade et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

counseling resources bar sharing of high-resolution genetic data with third parties in most countries. From both a medical and a scientific point of view, this "locking" of data is hardly compatible with good professional practice. For instance, for a physician or geneticist it may be essential to know whether a particular mutation found in the genome of their patient has been found in another patient with a similar phenotype before. Related questions are also likely to arise in basic research projects on both monogenic and complex (i.e. oliogenetic) diseases.

Tool description

We developed GrabBlur, a tool to collect and aggregate (i.e. "grab" and "blur") 'single nucleotide variants' (SNVs) linked to a specific trait or phenotype, and to share them with others by way of a public database while keeping individual samples unidentifiable. The database will not only help human geneticists to distinguish between benign variant findings and truly disease-causing mutations, but will also benefit genetic epidemiological research (i.e. case-control association studies) based upon large-scale SNV data.

In contrast to databases like ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) or the Human Gene Mutation Database (HGMD) [5], which only contain out-of-context information on genotype-phenotype associations, GrabBlur provides access to all SNVs detected in a given patient alongside the description of their specific phenotype. The Exome Variant Server (EVS) [6] provides about 2 million annotated SNVs of 6,500 individuals with heart-, lung- and blood-related diseases; more details are not specified. Through the straightforward aggregation of SNVs, it is not possible to find out, which SNV originated from which individual and phenotype. The EVS helps researchers excluding SNV candidates found in patients with monogenetic diseases, but it is not a resource to exchange genotypic and phenotypic data from other data sets, especially for Mendelian diseases where often the exact phenotypes are needed. Owing to this level of comprehensiveness, GrabBlur helps users not only to reckon known mutations, but also to validate newly found ones.

The most important feature of GrabBlur is the high level of anonymity ensured by its process of data aggregation. No conclusions as to the identity of a patient can be drawn even if the entire data stored for that individual are downloaded or the whole database is mirrored. It is possible neither to reconstruct a single patient genome nor to re-identify a patient from knowing their SNVs. Data is aggregated at the site of the submitter, i.e. behind their own firewall and under their responsibility for data protection. Hence, no identifying data leaves the submitter institution, and even if the data is "tapped" by an unauthorized person during upload to the database, a high level of privacy protection is maintained.

DNA sequence data are accepted by GrabBlur in standardized VCF format [7]. Additional information such as the phenotype or gender of a patient is stored in a separate "initialization file" (INI file format). Most of this information is optional and provision is at the discretion of the submitter. The following information may be recorded:

- **Trait**. A description of the disease of all patients in a GrabBlur set of samples (see below). Samples must be marked at least as 'patient' or 'healthy control'. (mandatory)
- **Phenotype**: GrabBlur uses Human Phenotype Ontology (HPO) terms [8] to classify phenotypes. Every phenotype can be ascribed an unlimited number of HPO terms. (optional)
- **Gender**: Gender of a single patient. (optional)
- **Platform**: DNA sequencing technology used. (optional)
- **Enrichment**: DNA enrichment kit used for sequencing. (optional)
- **PI**: Identity of principal investigator. (optional)
- **Contact details**: Identity, affiliation and e-mail address of the submitter (mandatory for upload, but optional release to public database)

To help users with the creation of the initialization file, we developed a web interface (Figure 1) to comfortably enter the required information, including sample ID and phenotype description. Although GrabBlur encodes phenotypes by a combination of HPO terms, users do not have to translate symptoms into numeric IDs. Instead, we employ an auto-completion procedure that finds all HPO terms matching the user input. The chosen terms are then presented in a tree structure, with their definitions accompanied by parent and children terms. This allows users to easily refine their description by choosing a more eligible term. In addition to marking symptoms as present, users can also identify particular symptoms as being absent to accentuate interesting characteristics of their patient.

On the project homepage, we also provide Perl scripts to read either a single VCF file or all VCF files contained in one directory and to directly submit filenames and sample IDs to this interface.

GrabBlur aggregates data in the following three steps:

1. Inspection of the additional information available for every patient

To prevent identification of a patient via the combination of different individual-specific informational items, these items must not be unique in the set of sample data provided to a third party. Every variant of a patient is associated with his meta-data. In case of uniqueness, the reconstruction of a patient's genome would be possible. At

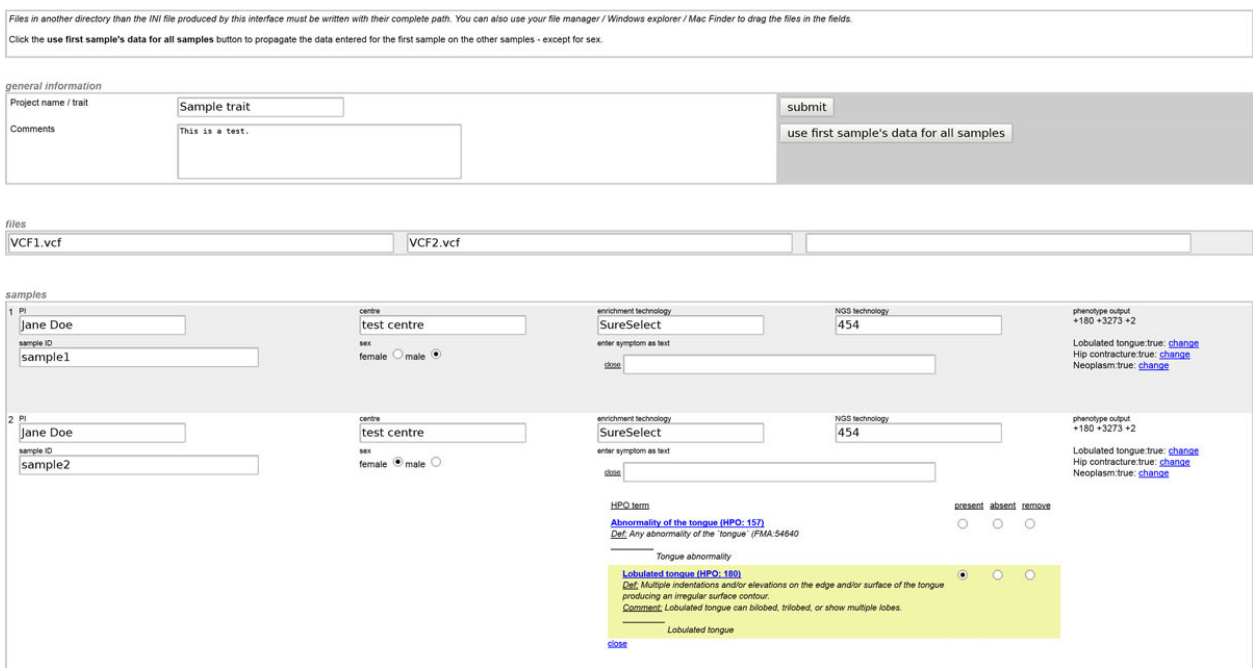


Figure 1 Creation of the initialization file. Screenshot of the interface to create the initialization files. The auto-completion procedure finds all HPO terms matching the user input. The chosen terms are then presented in a tree structure, with their definitions accompanied by parent and children terms.

least two samples must have exactly the same phenotypes, same gender information etc.

In order to generate a sufficient level of ambiguity, samples with an identical set of HPO terms are combined in classes. If any other additional information is not sufficiently ambiguous, GrabBlur blurs it by deleting e.g. the gender or the platform-name.

2. Fragmentation of the SNV-data

In a second step, the SNVs of a sample are divided into sub-samples of different size. A list linking sample IDs and sub-sample IDs is stored in a encrypted and password-protected file at the submitter site. Encryption is accomplished by means of the Blowfish algorithm of OpenSSL [9]. Only the submitter themselves can open this file. This is needed, for example, to delete a sample from the database in case the patient withdraws the consent.

Each SNV of a sample is randomly assigned to a sub-sample. This assignment is not uniformly distributed because otherwise any group of linked sub-samples would contain an approximately equal number of SNVs, thereby allowing reconstruction of the complete sample. Therefore SNVs are assigned to a sub-sample with a differently weighted likelihood.

3. Blurring the genotype information for rare variants

In a third step, all rare variants of a sample are aggregated by replacing the sub-sample ID of a rare SNV by the contact information of the submitting institution. Since a patient can easily be identified by singletons (i.e. SNVs that have been detected only once), these and other rare SNVs in their exome are blurred. In the aggregation step the association between an SNV and all belonging sub-samples has been deleted. Only the trait and (if known) the submitting institution remain linked to the SNV. Hence, only common SNVs carry a sub-sample ID and, therefore, are associated with specific phenotype information.

The threshold for a variant to be considered rare is variable and depends upon the submitted data. It is calculated from the median of all SNV frequencies as

$$freq(SNV) =< 1.5 * med(frq)$$

Here, *freq(SNV)* denotes the frequency of the SNV irrespective of its genotype, and *med(frq)* is the median over all SNV frequencies in the sample set. We choose the median because it is robust against outliers, like in this case above-average number of singletons.

The default factor of 1.5 can be modified by the submitter to get a lower or higher aggregation level. Usually, with a default factor of 1.5, the threshold equals between 8 and 12 so that a data set must comprise at least 8 to 12 samples in order to provide additional information other than the contact address.

Figure 2 shows the ratio of unblurred data (SNVs where the genotype is not aggregated) in relationship to the amount of aggregated samples using the default median factor of 1.5. The part of unblurred data increases logarithmically to the number of aggregated samples. Through the blurring of the rare variants the portion of unblurred data is reaching a plateau of about 70%. Hence, a good minimum sample size is n = 100 so that as less genotype information as possible gets blurred.

Figure 3 shows the frequency distribution of a sample set containing 50 randomly chosen exomes. As expected, the proportion of singletons is 10 times higher than that of common SNVs. In the given example, the threshold frequency for rare variants would be approximately 20% (or 10 occurrences).

Aggregation quality

To assess the aggregation quality and hence the level of ensured anonymity, a sample set of 10 individuals was blurred and compared to one randomly selected and non-aggregated sample of that set. For illustration, the sub-samples in Figure 4 were sorted according to the original sample IDs (sample ID, sub-sample ID on the X-axis). The selected individual is sample no. 6. It turns out that the overlap between a sample and its matching sub-samples is not notably larger than with other sub-samples. This is important if the data upload is intercepted or if the database itself gets compromised. Moreover, the aggregation also makes it impossible for an interested authority to identify an individual by comparing their own genetic data to the GrabBlur database (e.g. a law enforcement authority searching for a suspect).

Data access

After the aggregation steps, the blurred data is written into a new VCF file at the submitter site (Figure 5) from where they are being uploaded to the public database. This process does not start automatically so that the submitter keeps control of the data they provide. After uploading, other registered users are able to retrieve information on the submitted SNVs and their associated phenotypes. To access the GrabBlur database, we developed a web front end (accessible at http://grabblur.ikmb.uni-kiel.de) that offers two main features:

(1) After registration, users can upload their data. The web front end allows the user to choose an aggregated VCF-file, which must have been created before using the blurring software described above. The front end sends the file to a client software running on the same server, which checks the file for consistency and potential corruptions and then transfers it to the database.

During the upload-process, every SNV is automatically functionally annotated using our in-house software tool *snp*Acts (http://snpacts.ikmb.uni-kiel.de). *snp*Acts identifies whether an SNV causes a protein coding substitution and which amino acid is affected using the gene annotations

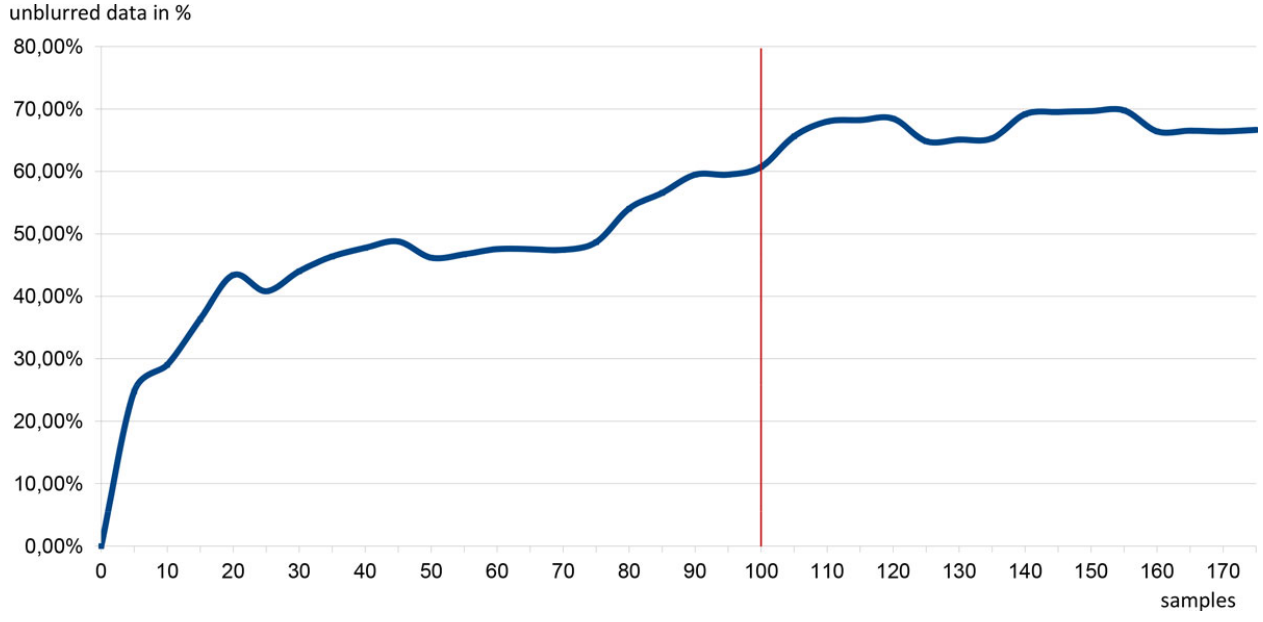


Figure 2 Ratio of the unblurred data with various sample set sizes. The ratio of unblurred data (SNVs where the genotype is not aggregated) in relationship to the amount of aggregated samples using the default median factor of 1.5. The part of unblurred data increases logarithmically to the number of aggregated samples. Through the blurring of the rare variants the portion of unblurred data is reaching a plateau of about 70%. Hence, a good minimum sample size is n = 100 so that as less genotype information as possible gets blurred. For blurred SNVs no genotype will be notified, only the contact data of the submitter will be named.

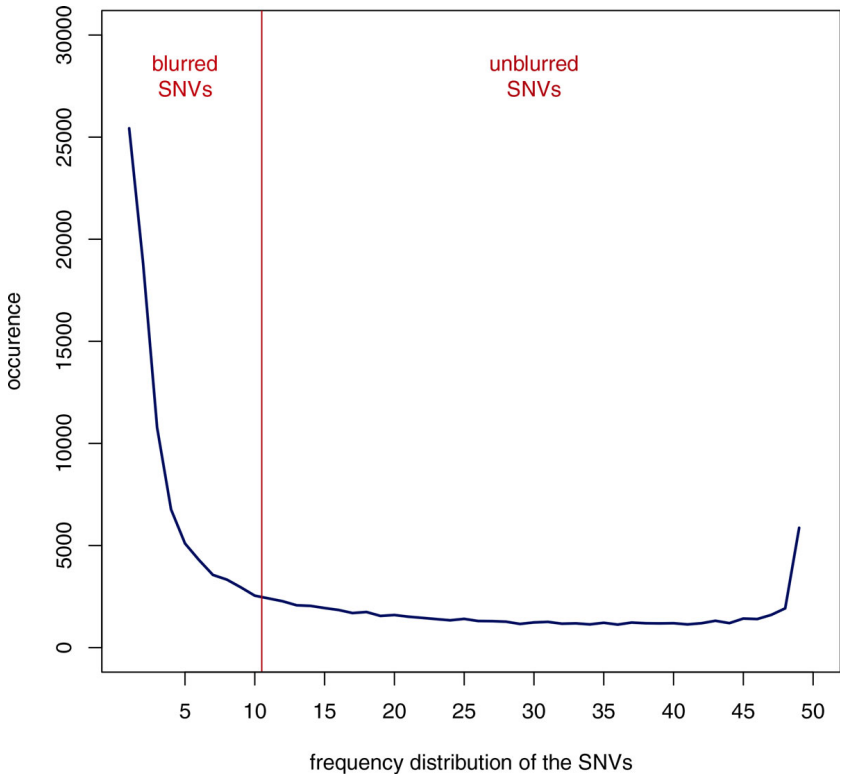


Figure 3 SNV frequency distribution in a sample of 50 exomes. Evidently, the proportion of singletons is 10 times higher than that of common SNVs. In the depicted example, the threshold frequency for rare variants was approximately 20% resp. 10 SNVs.

from CCDS [10] and RefSeq [11]. The amino acid changes in all iso-forms of the affected gene are classified and ranked in the following order: "nonsense" (most likely to be damaging), "readthrough", "start-lost", "splice site", "missense", "synonymous" (least likely to be damaging). To obtain more information for estimating whether an SNV is likely to be damaging, *snp*Acts also queries the Human Gene Mutation Database "HGMD" [5]. HGMD provides a database of comprehensive, in part manually curated data on human inherited disease mutations. Since this is a commercially available database, only an identifier from the HGMD database is named in *snp*Acts. All results of these annotations, including the highest ranked classification of the SNV, are stored in the database upon upload of the data.

(2) All registered users are able to search the database for loci of interest using either the chromosomal position, the dbSNP IDs of known SNVs, gene symbols, or a protein position in combination with a gene symbol. The latter is particularly useful to identify potentially compound heterozygote samples. However, phase information needs to be retrieved from the submitter via re-contact. Registered users can also perform combined searches simultaneously looking for terms of different type (e.g. "chr1:13272" and "rs6605067" and "NOD2"), so the search result contains information about the locus, i.e. weather it is situated within a gene, the gene identifier and gene function, and how many samples in the database carry an SNV at this locus (Figure 6). The allele and genotype frequencies of every SNV over all samples in the database are displayed as well as publicly available allele frequencies from the 1000 Genomes Project [12] (phase1) and from the Exome Variant Project [6] (ESP6500SI-V2). The user can access further information for each of these sets of samples, if provided by the submitter, including the associated trait in the form of HPO terms (Figure 7). Additional information, like the submitter contact information or the sample gender (Figure 8), can be obtained also if provided.

Implementation

The aggregation software was written in C++ on an Ubuntu Linux system. The runtime of the aggregation increases linearly with the amount of samples. The consumption of memory (RAM) increases logarithmically. On a desktop PC, a VCF file with 43,000 SNV was aggregated in less than 3 seconds using one core (Intel Xeon 4C, 2.0GHz). The aggregation of 50 exomes with about 40,000 - 45,000 SNVs needs approximately 128 MB RAM and 130 sec. The aggregation of 150 exomes needs about 7 minutes with approximately 350 MB RAM.

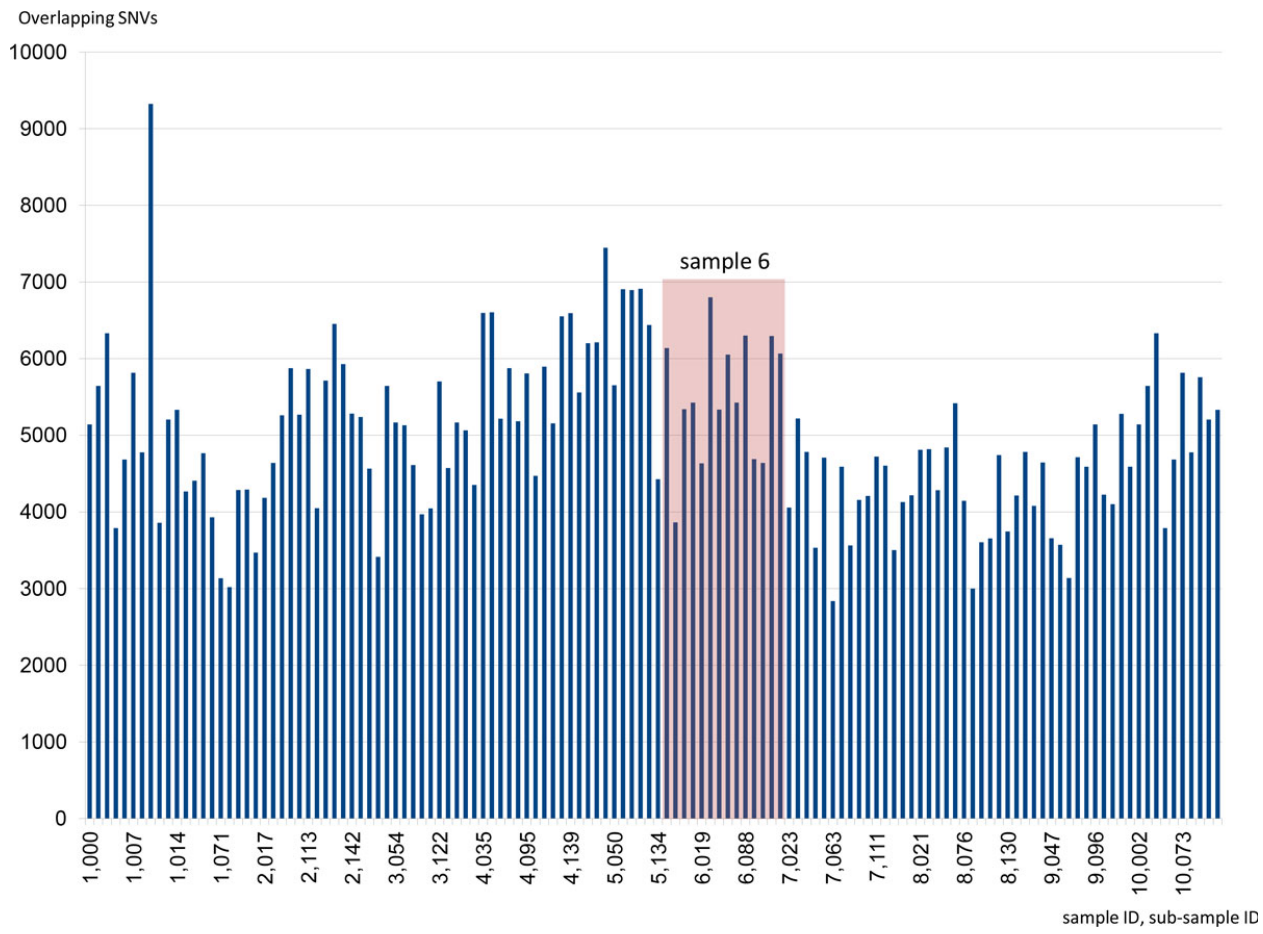


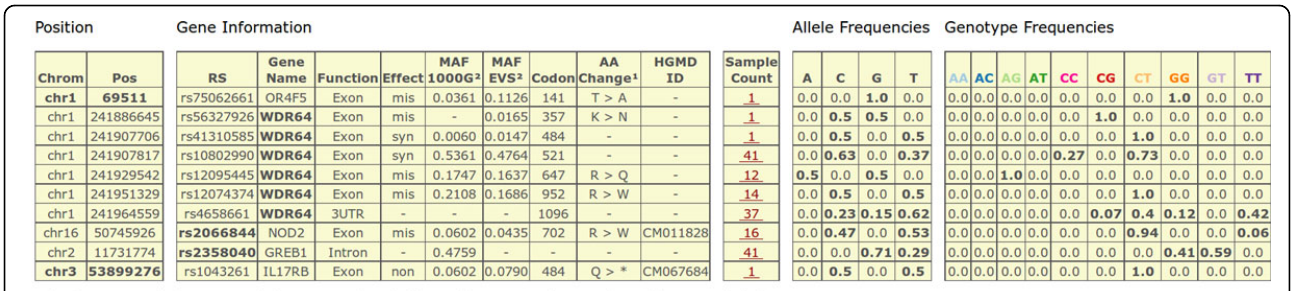
Figure 4 Assessment of the aggregation quality. To assess the aggregation quality of GrabBlur, a sample set of 10 individuals was "blurred" and compared to a non-aggregated sample of that set. For illustration purposes, the sub-samples were sorted according to the original sample IDs (sample ID, sub-sample ID on the X-axis). The corresponding individual is sample no. 6. The overlap between the sub-samples originating from this sample is not significantly higher than with the other samples.

The interface for the creation of the initialization files is programmed in Perl and uses JavaScript and AJAX to display the HPO terms retrieved from a PostgreSQL database.

The web interface for data access has been implemented using the Django web application framework [13] (v1.5.4) and the Python programming language [14] (v3.2.2). It is currently running on an Ubuntu Linux

```
##fileformat=VCFv4.1 (modified for presentation)
##SAMPLE=<ID=01, Contact=ikmb, Gender=male, Trait=IBD,HP=0002014,0000153>
##SAMPLE=<ID=02, Contact=ikmb, Gender=female, Trait=IBD, HP=0002014,0002027>
##SAMPLE=<ID=03, Contact=ikmb, Trait=IBD>
…
##CONTACT=<ID=ikmb, ShowContact=true>
…
#CHROM  POS   REF  ALT  INFO
chr1 1234 A  G  NS=100;AC=70;AF=0.35;GC_AA=50;GC_AG=30;GC_GG=20;
          SID_AA=01,02,03,…;SID_GC=11,12,13,…;SID_GG=23,24,25,…
chr2 2345 C  A,T  NS=100; AC= 71,14; AF=0.36,0.07;
          GC_AA=20;GC_AC=30;GC_AT=1;GC_CC=40;GC_CT=5;GC_TT=4;
          SID_AA=05,11,22,33,…;SID_AC=23,43,65,…;SID_CC=01,44,…;
          SID_AT=ikmb;SID_CT=ikmb;SID_TT=ikmb
```

Figure 5 Output file "aggregated VCF". GrabBlur writes the aggregated SNV data of all samples into a new VCF file. This figure shows a typical GrabBlur output of blurred SNVs. Some VCF information has been excluded, such as the dbSNP-ID and quality information, for explanatory purposes.



Position | Gene Information | Allele Frequencies | Genotype Frequencies

Chrom	Pos	RS	Gene Name	Function	Effect	MAF 1000G[2]	MAF EVS[2]	Codon	AA Change[1]	HGMD ID	Sample Count	A	C	G	T	AA	AC	AG	AT	CC	CG	CT	GG	GT	TT
chr1	**69511**	rs75062661	OR4F5	Exon	mis	0.0361	0.1126	141	T > A	-	1	0.0	0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0
chr1	241886645	rs56327926	**WDR64**	Exon	mis	-	0.0165	357	K > N	-	1	0.0	**0.5**	**0.5**	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0	0.0
chr1	241907706	rs41310585	**WDR64**	Exon	syn	0.0060	0.0147	484	-	-	1	0.0	**0.5**	0.0	**0.5**	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0
chr1	241907817	rs10802990	**WDR64**	Exon	syn	0.5361	0.4764	521	-	-	41	0.0	**0.63**	0.0	**0.37**	0.0	0.0	0.0	0.0	**0.27**	0.0	**0.73**	0.0	0.0	0.0
chr1	241929542	rs12095445	**WDR64**	Exon	mis	0.1747	0.1637	647	R > Q	-	12	**0.5**	0.0	**0.5**	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
chr1	241951329	rs12074374	**WDR64**	Exon	mis	0.2108	0.1686	952	R > W	-	14	0.0	**0.5**	0.0	**0.5**	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0
chr1	241964559	rs4658661	**WDR64**	3UTR	-	-	-	1096	-	-	37	0.0	**0.23**	**0.15**	**0.62**	0.0	0.0	0.0	0.0	0.0	**0.07**	**0.4**	**0.12**	0.0	**0.42**
chr16	50745926	**rs2066844**	NOD2	Exon	mis	0.0602	0.0435	702	R > W	CM011828	16	0.0	**0.47**	0.0	**0.53**	0.0	0.0	0.0	0.0	0.0	0.0	**0.94**	0.0	0.0	**0.06**
chr2	11731774	**rs2358040**	GREB1	Intron	-	0.4759	-	-	-	-	41	0.0	0.0	**0.71**	**0.29**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.41**	**0.59**	0.0
chr3	**53899276**	rs1043261	IL17RB	Exon	non	0.0602	0.0790	484	Q > *	CM067684	1	0.0	**0.5**	0.0	**0.5**	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0

[1] Only the amino acid substitution with the worst predicted effect on the protein is shown. Other isoforms may be lacking.
[2] All frequencies refer to the central European (CEU) ethnicity.

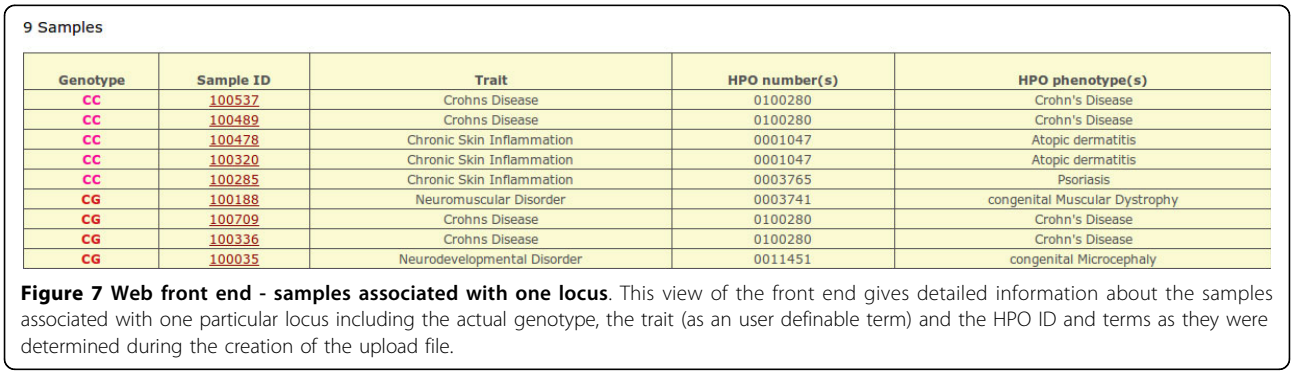
Figure 6 Web front end - search results. The figure shows an example search result for selected loci. The latter does not lie within a gene, so no additional gene information is provided. For the other two search terms the in-silico predicted gene function and effect of the SNV on the gene are provided. For all loci the number of samples in database with known genotype information is given - along with the frequencies for the genotypes and just the alleles.

Server (12.04.3 LTS). The MySQL database containing the actual GrabBlur data is located on another server (with the same configuration) and is accessed using the respective built-in modules of Django and Python.

Discussion

GrabBlur is a "light weight" tool to aggregate SNV data of thousands of samples with a specific trait or phenotype and to share the data with other via a public database. The main goal of GrabBlur, namely to keep each individual sample unidentifiable, was achieved by deleting other important information from individual exomes or genomes. For instance, all information of linked SNVs must be dropped to avoid the reconstruction of a given data set. But exactly this information is very valuable for scientific studies. For example, rare variant association analysis methods collapse rare variants into groups based upon, for example, the functional annotation of genomic regions. Whether GrabBlur can be used in such studies needs to be verified individually for each analysis method (for a review of methods, see [15]). However, GrabBlur is intended mainly to serve human geneticists who try to find more data on a variant and the phenotype of interest. The user-friendly GrabBlur web interface should inspire users to share their data and to use the tool for their own purposes. Although GrabBlur anonymizes the genetic data to a sufficient degree, a cautious user may want to use GrabBlur only behind their own firewall to handle aggregated information. While we encourage users to share their data, we also support such "internal" mirrors and provide instructions to set them up.

GrabBlur also has limitations that should not go unmentioned. For example, the system is not yet checking for duplicate uploads. It is thus possible that redundant data end up in the GrabBlur database. Moreover, the quality of an uploaded SNV may not have been adequately checked. Detailed quality data, as it can be generated using our previously reported tool pibase [16], would require that users also retrieve BAM files for their sequence data, run additional and standardized analyses. Moreover, the addition of the quality scores would significantly inflate the GrabBlur database. We rather prefer that submitter provide their contact details so that data users can enquire the quality of particular SNVs directly. The submitter may then go back to the raw data and use pibase, the Integrated Genomics Viewer [17] or other tools to assess the quality of the SNV in more detail. It is also possible with GrabBlur to ask submitters for additional details on the phenotype of a patient or for a detailed re-phenotyping based on new scientific findings.



9 Samples

Genotype	Sample ID	Trait	HPO number(s)	HPO phenotype(s)
CC	100537	Crohns Disease	0100280	Crohn's Disease
CC	100489	Crohns Disease	0100280	Crohn's Disease
CC	100478	Chronic Skin Inflammation	0001047	Atopic dermatitis
CC	100320	Chronic Skin Inflammation	0001047	Atopic dermatitis
CC	100285	Chronic Skin Inflammation	0003765	Psoriasis
CG	100188	Neuromuscular Disorder	0003741	congenital Muscular Dystrophy
CG	100709	Crohns Disease	0100280	Crohn's Disease
CG	100336	Crohns Disease	0100280	Crohn's Disease
CG	100035	Neurodevelopmental Disorder	0011451	congenital Microcephaly

Figure 7 Web front end - samples associated with one locus. This view of the front end gives detailed information about the samples associated with one particular locus including the actual genotype, the trait (as an user definable term) and the HPO ID and terms as they were determined during the creation of the upload file.

Sample information

Sample ID		1000306
Sex		female
Trait (as defined by uploader)		Crohn's Disease
HPO (Human Phenotype Ontology)	**Phenotypes**	Crohn's Disease
	Numbers	100280
Contact	**Name**	Andre Franke
	Institute	Institute of Clinical Molecular Biology
	Email	a.franke@mucosa.de

Figure 8 Web front end - sample details. In addition to the sample list (Figure 7) this detailed view provides the sex of the sample and the contact information provided by the uploader to get more information on the sample.

Competing interests
The authors declare they have no conflict of interests in relation to this SNP-SIG issue article.

Acknowledgements
This project received infrastructure support from DFG Cluster of Excellence "Inflammation at Interfaces". We thank the TMF working group "Molecular Medicine", especially Prof. Thomas Wienker, for critical discussions on and support of the project.

Declarations
The publication costs for this article were funded by the DFG Cluster of Excellence No. 306 „Inflammation at Interfaces".
This article has been published as part of *BMC Genomics* Volume 15 Supplement 4, 2014: SNP-SIG 2013: Identification and annotation of genetic variants in the context of structure, function, and disease. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S4

Authors' details
[1]Institute of Clinical Molecular Biology, Christian-Albrechts University of Kiel, Schittenhelmstraße 12, 24105 Kiel, Germany. [2]NeuroCure Clinical Research Centre, Charité - Universitätsmedizin, Charitéplatz 1, 10117 Berlin, Germany. [3]Institute of Medical Informatics and Statistics, Christian-Albrechts University of Kiel, Brunswiker Straße 10, 24105 Kiel, Germany. [4]TMF - Technology and Methods Platform for Networked Medical Research, Charlottenstraße 42, 10117 Berlin, Germany.

Published: 20 May 2014

References

1. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N: **What can exome sequencing do for you?** *J Med Genet* 2011, **48**:580-9.
2. **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).** [http://www.genome.gov/sequencingcosts].
3. Neveling K, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg E-J, Mensenkamp AR, Rodenburg RJT, Yntema HG, Spruijt L, Vermeer S, Rinne T, van Gassen KL, Bodmer D, Lugtenberg D, de Reuver R, Buijsman W, Derks RC, Wieskamp N, van den Heuvel B, Ligtenberg MJL, Kremer H, Koolen Da, van de Warrenburg BPC, Cremers FPM, Marcelis CLM, Smeitink JaM, Wortmann SB, van Zelst-Stams WaG, Veltman Ja, Brunner HG, *et al*: **A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases.** *Hum Mutat* 2013, (Denovo 281964).
4. Gilissen C, Hoischen A, Brunner HG, Veltman JA: **Unlocking Mendelian disease using exome sequencing.** *Genome Biol* 2011, **12**:228.
5. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN: **The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.** *Hum Genet* 2013.
6. **Exome Variant Server.** [http://evs.gs.washington.edu/EVS/].
7. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**:2156-8.
8. Robinson PN, Mundlos S: **The human phenotype ontology.** *Clin Genet* 2010, **77**:525-34.
9. **Blowfish.** [https://www.openssl.org/docs/crypto/blowfish.html].
10. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner M-M, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, *et al*: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.** *Genome Res* 2009, **19**:1316-23.
11. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33(Database)**:D501-4.
12. **The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061-73.
13. Django: **A Web framework for the Python programming language.** Django Software Foundation, Lawrence, Kansas, USA;[http://www.djangoproject.com].
14. **The Python Language Reference.** [http://docs.python.org/py3k/reference/index.html].
15. Bansal V, Libiger O, Torkamani A, Schork NJ: **Statistical analysis strategies for association studies involving rare variants.** *Nat Rev Genet* 2010, **11**:773-85.
16. Forster M, Forster P, Elsharawy A, Hemmrich G, Kreck B, Wittig M, Thomsen I, Stade B, Barann M, Ellinghaus D, Petersen B-S, May S, Melum E, Schilhabel MB, Keller A, Schreiber S, Rosenstiel P, Franke A: **From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software.** *Nucleic Acids Res* 2013, **41**:e16.
17. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29**:24-6.

doi:10.1186/1471-2164-15-S4-S8
Cite this article as: Stade *et al.*: GrabBlur - a framework to facilitate the secure exchange of whole-exome and -genome SNV data using VCF files. *BMC Genomics* 2014 **15**(Suppl 4):S8.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central