DOI: 10.14529/jsfi170201

Using High Performance Computing to Create and Freely Distribute the South Asian Genomic Database, Necessary for Precision Medicine in this Population

Asmi H. Shah[1], *Jonathan D. Picker*[1], *Saumya S. Jamuar*[1]

Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. Efforts to implement precision medicine have gained traction in recent years due to significantly increased understanding of the role of genetic variations in human disease over the past decade. However, delivery of precision medicine requires robust population specific reference genome datasets for full appreciation of existing natural variation. A large majority of publicly available genomic databases are primarily derived from Caucasian populations and do not fully address the diversity of Asian populations. In an effort to address this problem, we have aggregated and built a genomic database, ggcINDIA, specifically for South Asian populations. In collaboration with Global Alliance for Genomics and Health (GA4GH), we have made this database publicly available to the community through the GA4GH's Beacon project. ggcINDIA represents the first Beacon for South Asian populations. As more data are generated and aggregated, the ggcINDIA beacon will provide the precise genomic data that is critical to the delivery of precision medicine within South Asia.

Keywords: Precision Medicine, ggcINDIA, Beacon Network, South Asian Genome, Genome Data Sharing.

Introduction

Next generation sequencing and constant advances in the high throughput technologies as well as lab automation have made it possible to explore the vast variation that exists within the human genome [8, 14]. For example, variations in genes related to drug metabolism (also known as pharmacogenomics) such as *CYP2C19, NAT2,* etc. affect the individual's response to drug treatment. Similarly, presence of specific pathogenic variants in certain cancers allow the use of targeted therapeutics (also known as precision oncology). For example, treatment of melanoma with a somatic *V600E* variant in the *BRAF* gene, specifically includes the use of selective *BRAF* inhibitors such as *vemurafenib*. Selective inhibition of *BRAF* results in a relative reduction of 63% in risk of death and 74% in risk of tumor progression [2]. These success stories have accelerated the move towards precision medicine, a disruptive model of healthcare delivery, where treatment is tailored to the individual's characteristics, in most cases, the genetic or molecular information.

For successful delivery of precision medicine, it is imperative to understand the genomic variations, and their consequences, for different populations. Studies such as the 1000 Genome project have demonstrated that these genetic variations are dependent on the ancestry and ethnicity [6, 21]. They are responsible for the phenotypic diversity within the diverse populations, such as facial appearance, but more importantly, for differences in disease susceptibility and therapeutic response. A major limitation of previous genomic studies was a focus on Caucasian populations. More recent efforts have have begun to individuals from a more diverse non-Caucasian background, which has led to an increase in representation of non-European individuals in the NHGRI-EBI GWAS from 4% in 2009 to 19% in 2016 [18]. However, the vivid diversity of South Asian population [20], that is not accurately represented even with the increased representation

[1]Global Gene Corporation Pte Ltd, Singapore, Singapore

of non-European individuals in publicly available genomic databases [13, 18]. The concern with this bias is that it can result in misinterpretations and misdiagnoses [15]. In a recent analysis by the Exome Aggregation Consortium group at the Broad Institute, only 9 out of 192 variants, previously called as pathogenic, were truly pathogenic, while over 160 variants were population specific polymorphisms, and hence, likely benign [13, 23]. Furthermore, on comparing the standard datasets to bushmen in the KB1 African genome analysis, it was found that there was an increased frequency of sequence variation between them, and over 47% of variants identified were novel, affecting over 7700 genes; indicating the scale of population diversity [22].

To fill in the gap of South Asian genome knowledge, the goal of this study is to build genome database for South Asian populations and to make this database accessible to researchers and at large.

1. Methods

1.1. Aggregating Genomic Data of South Asian populations

Through separate projects, we identified individuals of South Asian ancestry, and performed genomic sequencing (either whole exome or whole genome) on these individuals. In addition, we aggregated available genomic data, including some available publicly through Creative Commons Attribution License [4]. All of the individuals included in this study had all of their four grandparents born on the Indian subcontinent.

The DNA sequence of each individual was collected in the standard FASTQ files, which store the DNA sequence as well as its corresponding quality scores.

1.2. Genome Analysis: Alignment, Variant Calling, and Variant Annotation

A customized bioinformatic analysis pipeline was set up to analyze the genome sequences and make the South Asian variant datasets available to the public (fig. 1).

The raw genomics data in FASTQ format were processed using the Sentieon DNAseq pipeline version 201611 [9]. Sentieon DNAseq is a proprietary reimplementation of Broad Institutes best practices pipeline for DNAseq [3] with an approximately 10x improvement in runtime. Performance improvements were achieved through use of Sentieon's proprietary improved algorithms and better resource management.

DNA sequences in FASTQ files were aligned to the known and publicly available reference genome GRCh37 using Sentieon BWA (sequence alignment tool) and the resulting alignments were sorted by genomic coordinates and converted to BAM format (binary format of sequence data) using the Sentieon UTIL binary. The quality of the aligned sequence data - including mean base quality for each flowcell cycle, the base quality score distribution, GC bias metrics, alignment metrics, and insert size metrics were calculated for each sample using the Sentieon driver. Duplicates in the sequences were removed, reads were realigned around indels (insertions and deletions in the sequence) identified by the 1000 Genomes project [6] or Mills et al. [17], and base quality scores were recalibrated. (The variation in the DNA sequence that occurs at a specific position in the genome is called a variant.) Variants were called for each sample independently using the Sentieon DNAseq Haplotyper and variants were output as genomic Variant Call Format (gVCF) files. Joint genotyping was performed on all gVCF files using the

Sentieon DNAseq GVCFtyper. This step creates a common VCF file having the information from all the individuals' sequences.

Variant Effect Predictor (VEP) version 86 from Ensembl was used to annotate the VCF for further analysis. VEP determines the effect of variants (including single nucleotide polymorphisms (SNPs), insertions and deletions) on genes, transcripts, and protein sequence, as well as regulatory regions [16].

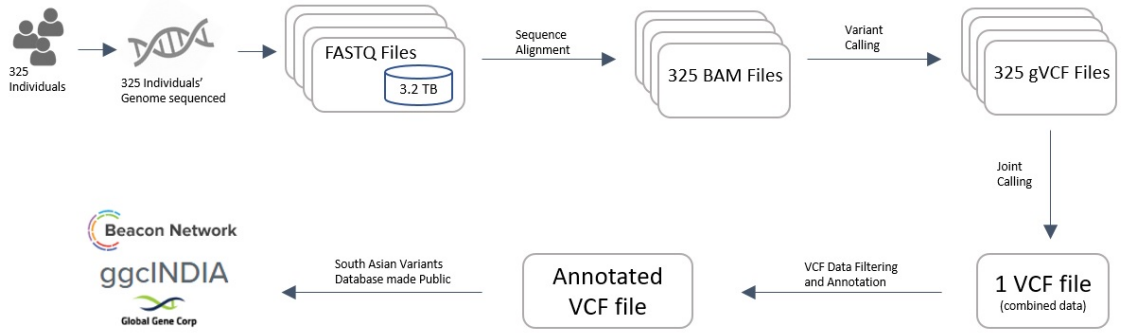

Figure 1. Genome Sequence Analysis of 325 individuals DNA sequences. From raw data genome files to making the Variants of South Asian populations public on Beacon Network search engine

1.3. Computation and Resources

High performance computing infrastructure provided through the National Super Computing Centre, Singapore was used to perform memory, resource, and compute intensive operations. PBS Pro was used to manage the workload and use the HPC resources efficiently. 2 units of 24 CPUs, 96GB of memory, 1 GPU, and 12 threads of 2 MPI processes were used. The total size of the raw data was 3183.3 GB.

Sentieon's DNAseq pipeline as well as VEP were deployed onto the compute nodes on the NSCC server infrastructure. Processing one individual's FASTQ files to a gVCF file took about 6 hours on a single 32 core server. 64GB of memory is recommended to process such a sample. The computation time can be reduced to under an hour using distributed computing processes on multiple parallel servers. In our case, with the above mentioned resource configuration, the job of processing 325 individuals' genome datasets was completed in 168 hours.

2. Results

2.1. Demographics

We aggregated genomic data from 325 individuals of South Asian ancestry (tab. 1). Out of the 234 individuals where data on geographical distribution within India was available, 67.2% were from North India, 15.9% were from South India, 14.7% were from West India and 2.2% were from East India. There were 291 (89.5%) males. The age range was 31 to 81 years (median 48 years).

Table 1. Distribution of 325 Individuals by their Country of Birth. For Each of them, all 4 grandparents were native to the Indian Subcontinent

COUNTRY OF BIRTH	PROPORTION(%)
India	72.0% (234)
Pakistan	10.5% (34)
Sri Lanka	5.2% (17)
Bangladesh	1.2% (4)
Afghanistan	0.6% (2)
Other	10.5% (34)

2.2. Genomic Data

All individuals underwent genomic sequencing as per standard protocols. 178 underwent whole genome sequencing and 147 underwent whole exome sequencing. All sequencing was performed on the Illumina platform.

2.3. Genomic Variant Calling and Annotation

Variants in one's genome are defined as the differences in an individual's genome when compared to a reference genome. These variants account for the differences among individuals and tend to cluster based on ancestry [7]. While some of these variations may directly alter the structure or function of the protein they code (also known as protein altering variants), a significant majority of these variants occur in the non protein coding regions, and the significance of these variants has not been well elucidated. Some of the variants could have associations with human diseases or complex traits.

In our cohort, we detected 19,643,311 variants, which were then annotated using VEP. The majority of the variants (81.6%) were single nucleotide variants (SNV) (fig. 2) (replacement of a single nucleotide in the sequence). The rest of the 18.4% variants are indels and sequence alteration - meaning there were insertions or deletions of nucleotide(s) from the sequence. Only 1.1% of these SNV variants were coding (coding region is the part which translates into proteins), while 47.1% were intronic and 39.7% were intergenic (fig. 3). Among the coding variants, 54.0% were missense variants, 42.0% were synonymous variants, 1.5% were frameshift variants and 1.0% affected the termination codon (fig. 4). Among the missense coding variants, 59.8% were predicted to be benign by Polyphen-2 [1], while 33.7% were predicted to be either possibly or probably damaging (suppl. fig. 1). Distribution of variants across chromosomes is demonstrated in suppl. fig. 2 to 27. In each of the figures, the X-axis is a position along the particular chromosome and the Y-axis is a number of variants at the given location. This distribution, in turn, does show the areas of increased variation.

2.4. ggcINDIA Beacon

Beacon Network by Global Alliance for Genomics and Health (GA4GH) is a global search engine for genetic mutations [10]. Each collaborator's genomic datasets in the form of VCF files are uploaded and is called lighting a 'Beacon'. It enables global discovery of genetic mutations,

Figure 2. Variant classification shown for the total of 19,643,311 variants detected among the 325 individuals. SNV= single nucleotide variant

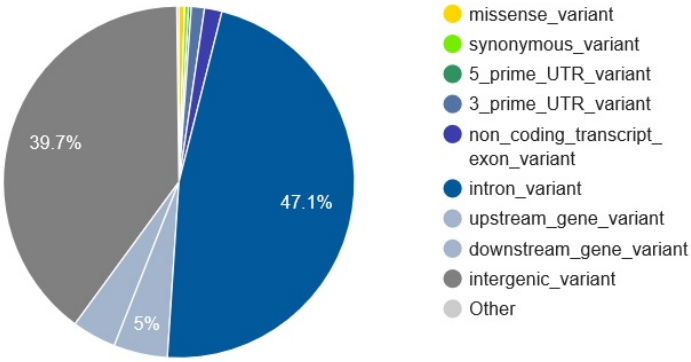

Figure 3. Variant classification depending on the consequences, shows the most severe ones. Only 1.1% are coding sequence variants

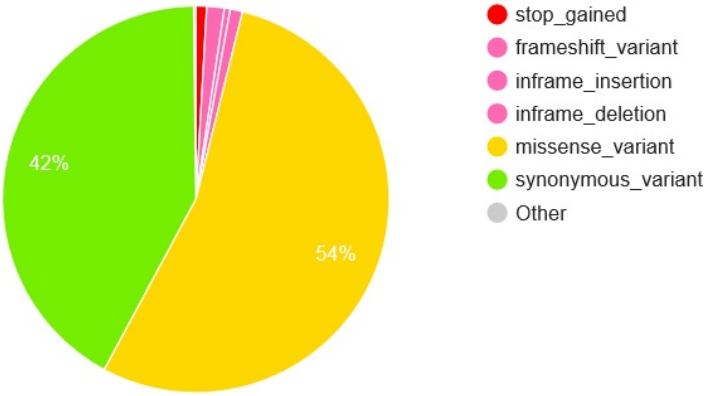

Figure 4. Coding consequences, variant classification of the 1.1% coding variants

federated across a large and growing network of shared genetic datasets. To join the effort and to contribute to the community, we chose Beacon Network as the platform to make our South Asian populations' variant database public.

In collaboration with GA4GH, we have published the South Asian genomic variant database, the first 'beacon' of its kind for the South Asian population called ggcINDIA (fig. 5). This beacon is the 69th beacon in the network. The beacon is a freely available resource and allows researchers and the public to query the presence or absence of a given variant detected in their own discovery cohort, and allows for filtering of variants for rarity. Once you access ggcINDIA, you can filter out the variants specifically within the South Asian population.

Over time, we foresee generating and aggregating more genome sequences of individuals from various cohorts of South Asian ethnicity into ggcINDIA beacon.

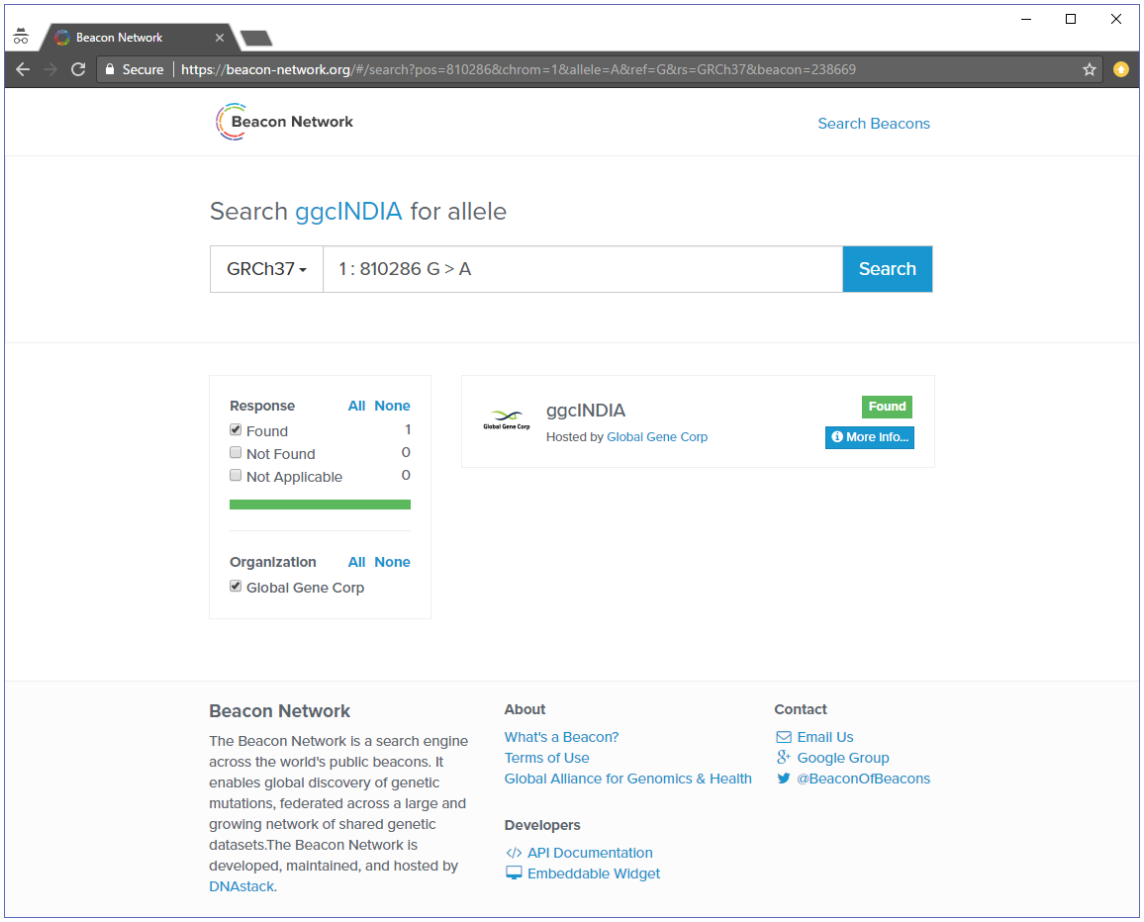

Figure 5. ggcINDIA on the Beacon Network. This interface allows researchers to know if a particular searched variant is present or absent in the population. Here, the searched variant was found in ggcINDIA.

3. Discussion

The correlation of genetic information with drug interactions as well as phenotypic and pathogenic traits has proven that healthcare can be improved by personalizing to ones characteristics and treatments [5]. Precision medicine is changing the dynamics of how healthcare is delivered. However, for precision medicine to have maximum impact, the genomics of diverse population cohorts must be known. The majority of known genetic knowledge is derived from Caucasian populations [18]. The relative frequency of alleles important for pharmacogenomics varies by population, meaning that certain drugs or drug groups will be less effective or even hazardous in some populations, *e.g.*, the risk for toxic epidermal necrolysis with the antiepileptic drug carbamazepine in East Asians [24], and more effective and safer in other such as statin use in Iranians with a specific *KIF6* variant [11].

ggcINDIA is an initiative that takes up the challenge to recruit the under-represented populations and add their genomic information to correct the known racial bias of currently available genomic knowledge. This study supports the fact that scientific data needs to be shared and made publicly available within the scientific community as well as the public [12, 19]. ggcINDIA

is part of global data sharing movement lead by GA4GH [19] and their flagship program of Beacon Network. Such initiatives will only widen the scope of the reference genome and take the necessary and obvious diversity into account.

ggcINDIA made its start with 325 individuals' genomic data. Our aim is to continue to grow and add data from more individuals to create a high fidelity South Asian reference genome. Thus, moving forwards, we invite other collaborators to come and share their genomic datasets for South Asian population and contribute in increasing the fidelity of the database. This process will provide more accurate genomic data that is critical to delivery of precision medicine within South Asia.

Acknowledgements

We would like to thank Sentieon Inc., USA for providing us access to the Sentieon DNAseq pipeline and Donald Freed from Sentieon Inc., USA for his assistance with the manuscript. We are also grateful for the assistance from the Singapore National Supercomputing Centre (NSCC), Global Alliance for Genomics and Health, European Bioinformatics Institute (EBI), Sanger Institute and DNAStack. Some of the data used in this study was made available through the Creative Commons Attribution License (© 2014 Chambers et al.).

References

1. Adzhubei, I., Jordan, D.M., Sunyaev, S.R.: Predicting functional effect of human missense mutations using polyphen-2. Current protocols in human genetics pp. 7–20 (2013), DOI:10.1002/0471142905.hg0720s76

2. Ascierto, P.A., Kirkwood, J.M., Grob, J.J., Simeone, E., Grimaldi, A.M., Maio, M., Palmieri, G., Testori, A., Marincola, F.M., Mozzillo, N.: The role of braf v600 mutation in melanoma. Journal of translational medicine 10(1), 85 (2012), DOI:10.1186/1479-5876-10-85

3. Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al.: From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics pp. 11–10 (2013), DOI:10.1002/0471250953.bi1110s43

4. Chambers, J.C., Abbott, J., Zhang, W., Turro, E., Scott, W.R., Tan, S.T., Afzal, U., Afaq, S., Loh, M., Lehne, B., et al.: The south asian genome. PLoS One 9(8), e102645 (2014), DOI:10.1371/journal.pone.0102645

5. Collins, F.S., Varmus, H.: A new initiative on precision medicine. New England Journal of Medicine 372(9), 793–795 (2015), DOI:10.1056/NEJMp1500523

6. Consortium, .G.P., et al.: A global reference for human genetic variation. Nature 526(7571), 68–74 (2015), DOI:10.1038/nature15393

[2] https://creativecommons.org/licenses/by-nc/3.0/

7. Cotton, R., Horaitis, O.: Human genome variation society. eLS (2006), DOI:10.1038/npg.els.0005964

8. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al.: A framework for variation discovery and genotyping using next-generation dna sequencing data. Nature genetics 43(5), 491–498 (2011), DOI:10.1038/ng.806

9. Freed, D.N., Aldana, R., Weber, J.A., Edwards, J.S.: The sentieon genomics tools-a fast and accurate solution to variant calling from next-generation sequence data. bioRxiv p. 115717 (2017), DOI:10.1101/115717

10. GlobalAllianceForGenomics&Health: Beacon network. `https://beacon-network.org`, accessed: 2017-05-09

11. Hamidizadeh, L., Abadi, B., Hosseini, R.H., Baigi, B., Ali, M., Dastsooz, H., Nejhad, A.K., Fardaei, M.: Impact of kif6 polymorphism rs20455 on coronary heart disease risk and effectiveness of statin therapy in 100 patients from southern iran. Archives of Iranian Medicine (AIM) 18(10) (2015)

12. Kosseim, P., Dove, E.S., Baggaley, C., Meslin, E.M., Cate, F.H., Kaye, J., Harris, J.R., Knoppers, B.M.: Building a data sharing model for global genomic research. Genome biology 15(8), 430 (2014), DOI:10.1186/s13059-014-0430-2

13. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., ODonnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.: Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616), 285–291 (2016), DOI:10.1038/nature19057

14. Levy, S.E., Myers, R.M.: Advancements in next-generation sequencing. Annual review of genomics and human genetics 17, 95–115 (2016), DOI:10.1146/annurev-genom-083115-022413

15. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., Kohane, I.S.: Genetic misdiagnoses and the potential for health disparities. New England Journal of Medicine 375(7), 655–665 (2016), DOI:10.1056/NEJMsa1507092

16. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., Cunningham, F.: The ensembl variant effect predictor. Genome biology 17(1), 122 (2016), DOI:10.1186/s13059-016-0974-4

17. Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., Devine, S.E.: An initial map of insertion and deletion (indel) variation in the human genome. Genome research 16(9), 1182–1190 (2006), DOI:10.1101/gr.4565806

18. Popejoy, A.B., Fullerton, S.M.: Genomics is failing on diversity. Nature 538(7624), 161 (2016), DOI:10.1038/538161a

19. Rahimzadeh, V., Dyke, S.O., Knoppers, B.M.: An international framework for data sharing: Moving forward with the global alliance for genomics and health. Biopreservation and biobanking 14(3), 256–259 (2016), DOI:10.1089/bio.2016.0005

20. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., Singh, L.: Reconstructing indian population history. Nature 461(7263), 489–494 (2009), DOI:10.1038/nature08365

21. Rotimi, C.N., Jorde, L.B.: Ancestry and disease in the age of genomic medicine. New England Journal of Medicine 363(16), 1551–1558 (2010), DOI:10.1056/NEJMra0911564

22. Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F., Qi, J., et al.: Complete khoisan and bantu genomes from southern africa. Nature 463(7283), 943–947 (2010), DOI:10.1038/nature08795

23. Song, W., Gardner, S.A., Hovhannisyan, H., Natalizio, A., Weymouth, K.S., Chen, W., Thibodeau, I., Bogdanova, E., Letovsky, S., Willis, A., et al.: Exploring the landscape of pathogenic genetic variation in the exac population database: insights of relevance to variant classification. Genetics in Medicine 18(8), 850–854 (2015), DOI:10.1038/gim.2015.180

24. Tangamornsuksan, W., Chaiyakunapruk, N., Somkrua, R., Lohitnavy, M., Tassaneeyakul, W.: Relationship between the hla-b* 1502 allele and carbamazepine-induced stevens-johnson syndrome and toxic epidermal necrolysis: a systematic review and meta-analysis. JAMA dermatology 149(9), 1025–1032 (2013), DOI:10.1001/jamadermatol.2013.4114