Errors in the interpretation of copy number variations due to the use of public databases as reference.

Nerea Bastida-Lertxundi[a§], Elixabet López-López[b§], M Angeles Piñán[c], Anna Puiggros[d], Aurora Navajas[e], Francesc Solé[d, f], Africa García-Orad[b*]

[a]Service of Biochemistry. University Hospital Álava. Osakidetza-Servicio Vasco de Salud, Vitoria, Spain.

[b]Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country (UPV/EHU), Leioa, Spain.

[c]Service of Hematology and Hemotherapy, University Hospital Cruces, Bilbao, Spain.

[d]Laboratori de Citogenètica Molecular. Servei de Patologia, Hospital del Mar, Barcelona, Spain.

[e]Unit of Pediatric Hematology/Oncology, University Hospital Cruces, Bilbao, Spain.

[f]Institut de Recerca contra la Leucèmia Josep Carreras. Badalona, Spain.

[§]Nerea Bastida-Lertxundi and Elixabet Lopez-Lopez contributed equally to this work.

Running title

CNV interpretation using public databases

Keywords

Acute lymphoblastic leukemia, deletions, duplications, DGV, germ-line sample

*Correspondence

Africa Garcia-Orad, Department of Genetics, Physic Anthropology and Animal Physiology, Faculty of Medicine and Dentistry-University of the Basque Country, Barrio Sarriena s/n, 48940 Leioa, Spain.

Phone: international +34.946012909. Fax: international +34.946013400

E-mail: africa.garciaorad@ehu.es

Abstract

The identification of new cryptic deletions and duplications can be used to improve prognostic classification in cancer. In order to obtain accurate results, it is necessary to discriminate between somatic alterations in the tumor cell and germ-line polymorphisms. For this purpose, copy number variations (CNVs) public databases have been used as reference. Nevertheless, the use of these databases may lead to erroneous results. Our main goal was to explore the limitations of the use of CNVs databases such as the Database of Genomic Variants (DGV) as reference. To that end, we used pediatric acute lymphoblastic leukemia (ALL) as a model.

We analyzed the genome-wide copy number profile of 23 ALL patients and conducted a comparison of the results obtained using DGV with those obtained using the normal sample from the patient as reference.

Using only the DGV, 19 % of alterations and 41 % of polymorphisms were erroneously catalogued.

Our results support the hypothesis that with the use of databases such as DGV as reference, a high percentage of the variations can be erroneously classified.

Introduction

The use of oligo-arrays helps identifying new cryptic alterations, such as deletions and duplications, which can be used to improve prognostic classification in cancer. However, it is known that some deletions and duplications are polymorphisms. Consequently, when we look for deletions and duplications in tumor cells, we need to discriminate the polymorphisms.

In order to identify the polymorphisms, databases such as the Database of Genomic Variants (DGV) have been used as reference in several studies [1]. DGV is a valuable web source that catalogues copy number variations (CNV) observed from studies of normal individuals. However, many CNVs may have only been observed in a single study, or with a single platform, and also may have not been validated by an alternative means [2]. Consequently, there is a possibility that a substantial amount of the catalogued data may be erroneous or incomplete [3-5]. Nevertheless, the errors in the interpretation of CNVs due to the use of the databases as reference have been poorly studied [6].

Our aim was to investigate the possible limitations of the use of CNV databases (using DGV as example) as reference to discriminate between polymorphisms and tumor alterations, using acute lymphoblastic leukemia (ALL) patients as a model.

Material and Methods

We compared the copy number alteration profile obtained from the tumor samples of 23 children, consecutively diagnosed with B-ALL at the University Hospital Cruces, using two different references for polymorphisms elimination. The references used were the DGV (2012 version) and the germ-line sample from the same patient.

Genomic DNA was obtained with QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) from lymphocytes isolated from bone marrow or peripheral blood with Ficoll-Paque™ PLUS

(GE Healthcare, Uppsala, Sweeden). Material collected at diagnosis (with more than 70% blast cells) was used as tumor sample; and in remission (with less than 5% blast cells) was used as paired germ-line sample. Informed consent was obtained from all patients or their parents before sample collection. The study was approved by the ethic committee of the University of the Basque Country (UPV/EHU). Copy number detection was carried out with the Cytogenetics Whole-Genome 2.7M platform (Affymetrix, Santa Clara, USA) and data analysis was performed with the Chromosome Analysis Suite (ChAS) software (Affymetrix, Santa Clara, USA).

Variations detected in both tumor and reference were considered polymorphisms and variations detected only in the tumor sample were considered alterations.

Results

We detected a total of 510 variations in the tumor samples, with an average of 22.2 genomic abnormalities per case, including aberrations and polymorphisms (Table S1).

In the first analysis conducted using the DGV as reference, 290 deletions/duplications were considered tumor alterations because they were not described in the database. The remaining 220 co-localized with variations described in the DGV and were considered as germ-line polymorphisms (Table 1).

In the second analysis performed with the remission sample from the same patient, 325 changes were classified as tumor alterations, as they were present only in the tumor sample. The other 185 were found in both samples from the same individual and were considered as germ-line polymorphisms (Table 1).

When we compared the results of both analyses, we observed that 56 (19%) variations, which were labeled as alterations because they were not included in the DGV, where detected in the germ-line sample and would have been catalogued as polymorphisms. And 91 (41%) variations, which were tagged as polymorphisms using the DGV as reference, where not detected in the germ-line sample and, then, would have been considered as alterations (Figure 1).

Discussion

The identification of new cryptic alterations, such as deletions and duplications, can be used to improve prognostic classification in cancer. In order to obtain good results, it is necessary to discriminate between somatic alterations that take place in the tumor cell and polymorphisms that appear in all the cells from the patient. For this purpose, different studies have used CNV public databases as reference [1]. Nevertheless, there is a possibility that a substantial amount of the catalogued data may be erroneous [4, 5]. However, there are few studies that explore the errors in the interpretation of CNVs due to the use of the databases as reference [6]. In this study our main goal was to explore the limitations of the use of CNV databases such as DGV as reference and to investigate whether normal material from the patient should be used for comparison. To that end, we used pediatric acute lymphoblastic leukemia as a model.

We searched for differences in results using the DGV or the normal sample from the same patient, as reference for polymorphisms discrimination. We detected 510 changes in the tumor tissue of 23 B-ALL patients. Using the DGV as reference, 290 were classified as tumor alterations and 220 as polymorphisms. By contrast, using the normal sample from the own patient as reference, 325 were classified as alterations and 185 as polymorphisms. Comparing both results, we could see that, using only DGV, we found erroneous results, in agreement with that observed in a population of patients with myelodisplastic Syndromes [6].

We found that, using DGV, we obtained 56 erroneous alterations that should have been classified as polymorphisms because they were found in the germ-line sample. This may happen due to the fact that CNV databases are far from complete. On the one hand, many polymorphisms do not appear frequently enough to be included in the database [4]. On the other hand, the studies compiled in the database are performed with platforms with different probe coverage and resolution [7] and, depending on the platform used in each case, some CNVs remain undetected. Consequently, many polymorphisms that are present in the patients are not collected in the databases, as we have confirmed in our population. We also considered the possibility that the persistence of abnormal ALL clones in the remission sample or the presence of circulating free DNA (cfDNA) from tumor cells could be confounding factors. However, it is unlikely since the percentage of aberrant DNA molecules that can be detected on Affymetrix Cytogenetics Whole-Genome 2.7M arrays is 20%, according to the manufacturer. In our study, the number of tumor cells in the remission sample was between 0 and 5% and the percentage of residual cfDNA from tumor cells remaining within the pellet of lymphocytes after gradient isolation and washing should be very low.

And, also, 91 of the alterations catalogued as polymorphisms, due to co-localization with polymorphisms described in the same frame in the DGV, where not observed in the normal tissue and would have been classified as alterations. Some of them (34/91, 37%) include T-cell receptor genes and IGL loci (7p14.1, 7q34, 14q11.2, 22q11.22). The detection of alterations in these loci is usually explained by the clonal origin of the tumor in those cases rather than to the leukemic process in itself. The rest of alterations (57/91, 63%) would have not been considered using only the DGV. This could be due to the fact that the database includes submissions that have not been thoroughly validated [4]. Many CNVs may have only been observed in a single study, or with a single platform, and also may have not been validated by an alternative means [2]. This raises the possibility that a substantial amount of the catalogued data may be erroneous and, as a result, we could find false polymorphisms co-localizing with alterations present in the

patients. Therefore, caution should be exercised when relying heavily on such databases [8]. In addition, recent studies based on very high-resolution arrays have shown that the size of CNV could be smaller than what was recorded in DGV, mainly because of a resolution bias [3]. Consequently, alterations that are next to a wrongly limited polymorphism could be considered as such. Moreover, it has been observed that regions that are frequently altered by CNV are more likely to undergo somatic copy number changes in stressed conditions because of their complex genomic architecture, they are prone to change [9]. Those changes that would have been considered polymorphisms and excluded using the DGV are alterations that could contribute to disease progression. Anyway, it should be taken into consideration that even if an alteration takes place in the tumor cell, it does not have to be pathogenic of necessity, and that it could be a benign, passenger mutation.

Consequently, we have shown that studies performed using the DGV as reference [1] could be providing some erroneous results.

In conclusion, the results of this study support the idea that the cytogenetic characterization of tumor tissue with databases such as DGV could give erroneous results. Consequently, in order to carry out a good characterization of the patient the paired normal tissue is required.

Acknowledgements

This project was supported by RTICC (RD/06/0020/0048; RD12/0036/0044), Basque Government (IT663-13, SAI10/03 and 2006111015), and UPV/EHU (UFI11/35). ELL was supported by a "Fellowship for recent Doctors until their integration in postdoctoral programs" by the Investigation Vice-rector's office of the UPV/EHU. Support by the Genomics Unit of the CIMA is gratefully acknowledged.

References

[1] Lundin C, Hjorth L, Behrendtz M, Nordgren A, Palmqvist L, Andersen MK, Biloglav A, Forestier E, Paulsson K, Johansson B. High frequency of BTG1 deletions in acute lymphoblastic leukemia in children with down syndrome. Genes Chromosomes Cancer 2012;51:196-206.

[2] Vermeesch JR, Brady PD, Sanlaville D, Kok K, Hastings RJ. Genome-wide arrays: quality criteria and platforms to be used in routine diagnostics. Hum Mutat 2012;33:906-15.

[3] Duclos A, Charbonnier F, Chambon P, Latouche JB, Blavier A, Redon R, Frébourg T, Flaman JM. Pitfalls in the use of DGV for CNV interpretation. Am J Med Genet A 2011;155A:2593-6.

[4] Heinrichs S, Li C, Look AT. SNP array analysis in hematologic malignancies: avoiding false discoveries. Blood 2010;115:4157-61.

[5] Ostrovnaya I, Nanjangud G, Olshen AB. A classification model for distinguishing copy number variants from cancer-related alterations. BMC Bioinformatics 2010;11:297.

[6] Heinrichs S, Kulkarni RV, Bueso-Ramos CE, Levine RL, Loh ML, Li C, Neuberg D, Kornblau SM, Issa JP, Gilliland DG, Garcia-Manero G, Kantarjian HM, Estey EH, Look AT. Accurate detection of uniparental disomy and microdeletions by SNP array analysis in myelodysplastic syndromes with normal cytogenetics. Leukemia 2009;23:1605-13.

[7] de Leeuw N, Dijkhuizen T, Hehir-Kwa JY, Carter NP, Feuk L, Firth HV, Kuhn RM, Ledbetter DH, Martin CL, van Ravenswaaij-Arts CM, Scherer SW, Shams S, Van Vooren S, Sijmons R, Swertz M, Hastings R. Diagnostic interpretation of array data using public databases and internet sources. Hum Mutat 2012.

[8] Lee C, Iafrate AJ, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet 2007;39:S48-54.

[9] Starczynowski DT, Vercauteren S, Sung S, Brooks-Wilson A, Lam WL, Karsan A. Copy number alterations at polymorphic loci may be acquired somatically in patients with myelodysplastic syndromes. Leuk Res 2011;35:444-7.

Table legends

Table 1. Summary of the classification of the alterations.

Tumor tissue vs. Dgv		**Tumor tissue vs. Normal tissue**
Classification	**Erroneously classified**	**Classification**
290 alterations	56 (19%)	325 alterations
220 polymorphisms	91 (41%)	185 polymorphisms

Figure legends

Figure 1. Differences in results observed using the DGV and the normal tissue. A) A variation that is classified as a deletion using the normal tissue as reference and that is classified as a polymorphism using the DGV as reference. B) A variation that is classified as polymorphism with the normal tissue and as deletion with the DGV (because no polymorphism is described in the region).