BMC Genomics

SOFTWARE

Open Access

CoNVaQ: a web tool for copy number variation-based association studies

Simon Jonas Larsen[1]*, Luisa Matos do Canto[4], Silvia Regina Rogatto[3] and Jan Baumbach[2,1]

Abstract

Background: Copy number variations (CNVs) are large segments of the genome that are duplicated or deleted. Structural variations in the genome have been linked to many complex diseases. Similar to how genome-wide association studies (GWAS) have helped discover single-nucleotide polymorphisms linked to disease phenotypes, the extension of GWAS to CNVs has aided the discovery of structural variants associated with human traits and diseases.

Results: We present CoNVaQ, an easy-to-use web-based tool for CNV-based association studies. The web service allows users to upload two sets of CNV segments and search for genomic regions where the occurrence of CNVs is significantly associated with the phenotype. CoNVaQ provides two models: a simple statistical model using Fisher's exact test and a novel query-based model matching regions to user-defined queries. For each region, the method computes a global q-value statistic by repeated permutation of samples among the populations. We demonstrate our platform by using it to analyze a data set of HPV-positive and HPV-negative penile cancer patients.

Conclusions: CoNVaQ provides a simple workflow for performing CNV-based association studies. It is made available as a web platform in order to provide a user-friendly workflow for biologists and clinicians to carry out CNV data analysis without installing any software. Through the web interface, users are also able to analyze their results to find overrepresented GO terms and pathways. In addition, our method is also available as a package for the R programming language. CoNVaQ is available at https://convaq.compbio.sdu.dk.

Keywords: Copy number variation, Association studies

Background

Copy number variation (CNV) is a type of structural variation in the genome in which a large segment of the DNA is either duplicated or deleted. Genome-wide association studies (GWAS) have been an important tool for discovering associations between genomic variants and disease phenotypes. GWAS data analysis methods have generally focused on single-nucleotide polymorphisms (SNPs) but can be applied to CNVs as well in order to determine the impact of larger structural variations on traits or phenotypes. Recent studies have shown that a large number of CNVs are present in healthy individuals, and are a significant source of genetic diversity in the population [1]. Currently, the Database of Genomic Variants reports more than half a million CNVs with most variations ranging from 1 kb to 10 kb in size [2] .

Copy number variations affecting individual genes have been linked to the susceptibility of HIV/AIDS [3], risk to develop psoriasis [4], and autism spectrum disorders [5], amongst others. CNVs have also been shown to influence gene expression [6, 7]. Functional genomic alterations may contribute to the development and progression of diseases [8]. Thus, measuring copy numbers of such genes alongside their expression may potentially also improve diagnostics.

Due to the high number of CNVs observed in healthy individuals and the rarity of disease-associated CNVs observed even in individuals with the disease, sophisticated methods are necessary in order to detect statistically significant CNVs. A large number of tools for calling CNVs from microarray and next-generation sequencing data currently exists [9, 10]. However, few methods are available for identifying CNVs associated with a pheno-

*Correspondence: sjlarsen@imada.sdu.dk
[1] Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, DK-5230, Odense, DK, Denmark
Full list of author information is available at the end of the article

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. 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.

type. CNVRuler [11] is a graphical desktop application that builds CNV regions using one of three models: overlapping regions, reciprocal overlap and segmentation at CNV boundaries. Associations between CNV regions and phenotype are determined using either Fisher's exact test, a chi-squared test, linear regression or logistic regression. ParseCNV [12] is a suite of command line tools for CNV-based association studies. It performs significance testing on probe-level using Fisher's exact test. Probes in close proximity with similar p-values are then merged into CNVRs. To our knowledge, no web service and no query-based methods exist yet.

Implementation

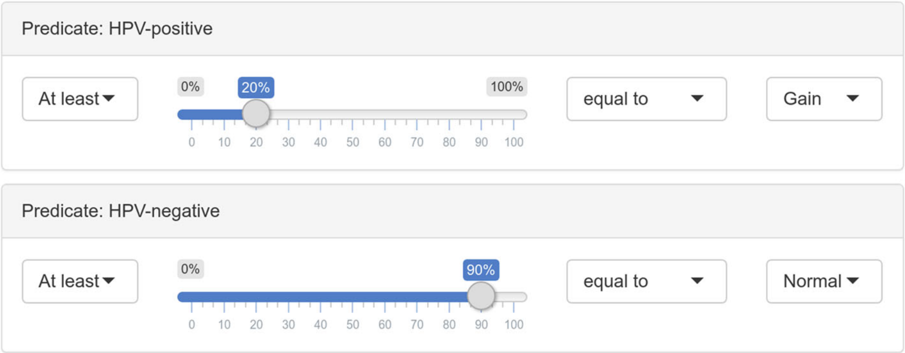
In this work we present CoNVAQ, a new web-based tool for copy number variation-based assocation study data analysis. Our method allows users to upload two sets of segmented CNVs (e.g. disease and healthy groups) and search for CNV regions where the occurence of CNVs is significantly associated with the classification of the samples (phenotype). Our software provides two models for signifiance testing. The first model is a traditional statistical model using Fisher's exact test for testing significance of associations between CNV and phenotype similar to what is implemented in previous methods. The second model is a novel query-based model, that allows users to specify what patterns are considered significant using simple queries through the web interface (Fig. 1). While not as statistically robust, the second model is able to capture patterns that may not show up using a statistical hypothesis test, and is, in our opinion, easier to understand and interpret. For each CNV region found, our method computes an empirical q-value by repeated permutation of the samples between the two groups, in order to estimate significance on a genome-wide scale. Users are able to inspect the individual reported regions to obtain a distribution of events and examine in which samples a variation is observed. Our web tool also provides a gene set enrichment analysis allowing users to search for an overrepresentation of Gene Ontology (GO) terms, KEGG and Reactome pathways or disease associations (Fig. 2) among the genes located in the discovered CNV regions. Our web server works on segmented CNV calls and does not produce CNV calls from raw data, which is left to one of the many existing tools. Hence, our tool can be applied to any CNV data set regardless of the technology used. CoNVaQ is an easy-to-use web tool, where all results are computed remotely on our servers, making it usable from any desktop PC with a web browser installed. Furthermore, we also provide CoNVaQ as a package for the R programming language, allowing researchers to run analyses locally.

Copy number variation region definition

We define a copy number variation region (CNVR) as a genomic region within a single chromosome wherein no sample changes state. Each chromosome is initially segmented into regions such that a new region starts at every end point (start or end) for every CNV among all samples (Fig. 3). As a result of this, no sample changes state within a region – only at region boundaries. Furthermore no two adjacent regions will be identical because at least one sample must change state in order for a new region to start, but two adjacent regions may have the same distribution of CNVs.

Statistical model

The statistical model uses Fisher's exact test for computing the significance of association between two groups of samples for some CNV event. The method considers each type of event (i.e. loss, gain and LOH) separately and classifies each sample as either having a variation of that type in the region or not. For each CNVR a 2 × 2 contingency table is built and a p-value is computed. Any region with



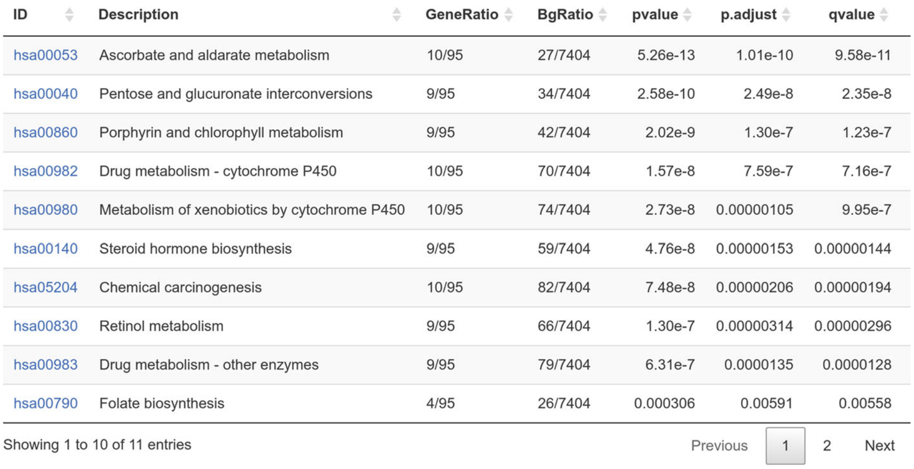
Predicate: HPV-positive

At least 20% equal to Gain

Predicate: HPV-negative

At least 90% equal to Normal

Fig. 1 Example of a query specified through the web interface for the query-based model. This query specifies that the method should look for regions where at least 20% of the samples in the positive group have a gain in copy numbers and at least 90% of the samples in the negative group have no variation



ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
hsa00053	Ascorbate and aldarate metabolism	10/95	27/7404	5.26e-13	1.01e-10	9.58e-11
hsa00040	Pentose and glucuronate interconversions	9/95	34/7404	2.58e-10	2.49e-8	2.35e-8
hsa00860	Porphyrin and chlorophyll metabolism	9/95	42/7404	2.02e-9	1.30e-7	1.23e-7
hsa00982	Drug metabolism - cytochrome P450	10/95	70/7404	1.57e-8	7.59e-7	7.16e-7
hsa00980	Metabolism of xenobiotics by cytochrome P450	10/95	74/7404	2.73e-8	0.00000105	9.95e-7
hsa00140	Steroid hormone biosynthesis	9/95	59/7404	4.76e-8	0.00000153	0.00000144
hsa05204	Chemical carcinogenesis	10/95	82/7404	7.48e-8	0.00000206	0.00000194
hsa00830	Retinol metabolism	9/95	66/7404	1.30e-7	0.00000314	0.00000296
hsa00983	Drug metabolism - other enzymes	9/95	79/7404	6.31e-7	0.0000135	0.0000128
hsa00790	Folate biosynthesis	4/95	26/7404	0.000306	0.00591	0.00558

Showing 1 to 10 of 11 entries

Fig. 2 Example of output from gene set enrichment. The table shows KEGG pathways that are significantly enriched with the genes from a set of CNV regions

a p-value less than the user-specified threshold will be reported along with the observed event.

Query-based model

The query-based model works by extracting all regions matching some user-specified query. The user must specify a predicate for the two groups of samples being compared. Then, regions where both groups match their respective predicate are identified and reported as part of the result.

A predicate is defined as a tuple (I, R, E, T), where $I \in \{\leq, \geq\}$, $R \in [0,1]$, $E \in \{=, \neq\}$ and $T \in \{\text{Normal}, \text{Gain}, \text{Loss}, \text{LOH}\}$. A query Q is defined by a pair of predicates $Q = (P_1, P_2)$. An example of such a query could be $P_1 = (\geq, 0.2, =, \text{Gain})$ and $P_2 = (\leq, 0.1, \neq, \text{Normal})$, describing that all regions in which at least 20% of samples in the first group have a gain in copy numbers while at most 10% of samples in the second group may have any kind of CNV, regardless of type, are considered significant. An example is illustrated in Fig. 4.

Q-value computation

Empirical q-values are computed for each reported CNVR by repeatedly perturbing the distribution of samples among the two populations. Samples are distributed among the two populations such that the original population sizes are preserved. For each of the found CNVRs we compute how often we see a CNVR that is equally or more significant in each of the repetitions when performing the same query. A region is considered more significant if it spans a larger number of base pairs. This is based on the following reasoning: The null hypothesis is that the occurrence of each CNV is independent of the phenotype of the sample. Under this hypothesis, larger regions of overlapping CNVs are less likely to occur by chance than smaller regions (under the assumption

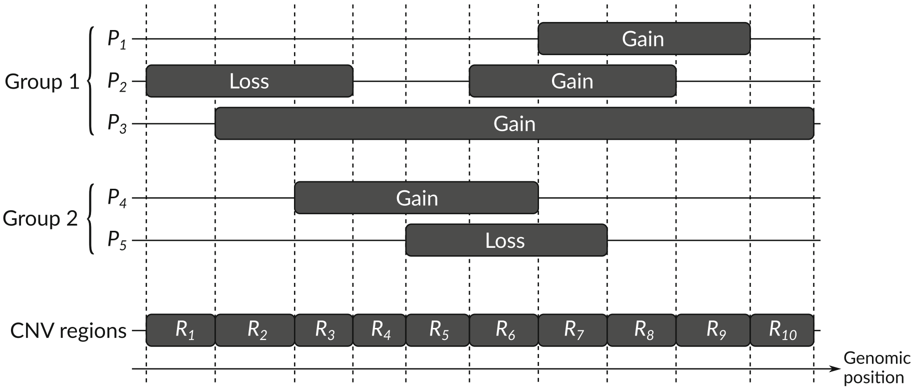


Fig. 3 Example of chromosome segmented into ten CNV regions by CNVs from five patients. A new region is started whenever a segment from any patient starts or ends

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
Group 1	Gain	0%	33%	33%	33%	33%	67%	100%	100%	67%	33%
	Loss	33%	33%	33%	0%	0%	0%	0%	0%	0%	0%
Group 2	Gain	0%	0%	33%	33%	33%	33%	0%	0%	0%	0%
	Loss	0%	0%	0%	0%	33%	33%	33%	0%	0%	0%
	P_1						✔	✔	✔	✔	
	P_2	✔	✔	✔	✔			✔	✔	✔	✔
	$P_1 \wedge P_2$							✔	✔	✔	

P_1 = (≥, 60%, =, Gain), P_2 = (≤, 40%, ≠, Normal)

Fig. 4 Example of matching query to CNV regions. The query $((\geq, 0.60, =, \text{gain}), (\leq, 0.40, \neq, \text{normal}))$ is evaluated against the ten CNV regions generated in Fig. 3. LOH frequencies have been left out for simplicity. Four regions match the predicate for group 1 (R_6-R_9), and eight regions match the predicate for group 2 (R_1-R_4, R_7-R_{10}). Three regions match the full query (R_7, R_8 and R_9)

that most of the genome does not exhibit any variation).

Only CNVRs of the same type are compared, i.e. finding a larger duplication will not affect the q-value of a deleted region. The position, length and type of each CNV is preserved under the perturbation. This preservation is important because we cannot reasonably assume the positions in which CNVs appear in the genome are random. By redistributing the samples among the populations while preserving their size, we instead compute the probability of observing a given overlap if there is no contingency between phenotype and the classification of samples.

Merging adjacent CNVRs

In some cases the CNV calling method might detect two or more CNVRs in very close proximity, separated only by a small number of base pairs. Such regions may correspond to just one region with some internal variation. Furthermore, when segmenting the genome into CNV regions as described above, we may produce several regions in a row with very similar variation distributions. In order to consider such regions as singular CNVRs, CoNVaQ includes an option to merge adjacent regions within some user-specified number of base pairs. After all matching CNVRs have been selected, adjacent CNVRs of the same type that are within this threshold will be merged into a single CNVR. Regions are merged before q-values are computed, and the merging step is also performed for each repetition when computing q-values as well. For the statistical model, the p-value of the new region will be the largest value (least significant) of regions being merged. Regions returned from the query-based model do not have a type, meaning they can be merged with any other region if within the threshold. The frequencies of merged regions will be represented as a range, e.g. if two regions with loss of copy numbers in 23 and 31% of samples, respectively, are merged, the new region will report loss in 23-31% of samples. The length of the new region will include the gap between merged regions as well. Note that merged regions do not match the previous definition of CNVRs as each sample is no longer guaranteed to have the same state for the entire span of the region.

Enrichment analysis

For the reported CNVRs, users are able to select one or more regions and extract all known genes overlapping those regions. A database of known genes was obtained from Ensembl [13]. A gene is said to be overlapping a CNVR if their genomic regions share at least one base pair. CoNVaQ also provides a gene set enrichment analysis. For the set of reported genes overlapping a CNVR, users can search for overrepresented Gene Ontology terms [14], KEGG pathways [15], Reactome pathways [16], Disease Ontology terms [17] and DisGeNET disease associations [18]. Enrichment analysis is carried out using the DOSE [19] and clusterProfiler [20] R packages. Statistical significance of enrichment is determined using a hypergeometric test. Let G be the set of genes overlapping the found CNV regions and C be the gene set we want to investigate for enrichment. Then a p-value is computed as

$$P(X \geq k) = \sum_{i=k}^{\min(K,n)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where N is the number of all genes, K is the number of genes in C, n is the number of genes in G and k is the number of genes both in G and C. Adjusted p-values are also computed using the Benjamini-Hochberg procedure, as well as estimated q-values using the method described in [21].

Software requirements

CoNVaQ is implemented as a web tool accessible through a web browser. All parsing of data and computation of results is done remotely on the server, and results are then displayed in the web interface. As such, only a modern, HTML5-enabled web browser supporting Javascript is necessary in order to use CoNVaQ.

Results

To demonstrate our platform we analyzed the penile cancer (PC) data set from [22]. It contains segmented CNV calls from 41 penile squamous cell carcinomas samples, where 14 samples were identified as HPV-positive and the remaining 27 as HPV-negative. We performed an association study between HPV-positive vs. HPV-negative samples in order to identify genomic variations that were more common in the HPV-positive group, using the two models implemented in CoNVaQ.

Statistical model

We first searched for significant CNVs using the statistical model. The statistical model uses Fisher's exact test to compute *p*-values for each individual CNV. Then, q-values are computed to estimate the probability of seeing a significant CNV of this type and size over the entire genome (statistical model and q-value computation are detailed in Methods section). We used a *p*-value cutoff of $\leq$ 0.05 for significance and enabled merging of adjacent CNVs with a distance threshold of 0 base pairs (i.e. only directly adjacent regions are merged). The method found 16 CNV regions (CNVRs) with significant *p*-value in chromosomes 2, 3, 4, 5, 8, 9, 16, 17 and 19 (Table 1). Q-values ranged from 0.0955 to 0.8415, meaning none of the regions had a statistically significant q-value (< 0.05). The most statistically significant region was a large loss event in chromosome 4. The 16 regions found here were also previously reported in [22].

Query-based model

We next also searched for significant CNVRs using the query-based model. The query-based model finds regions matching some user-specified query, and q-values are then computed using the same procedure as for the statistical model (detailed in Methods section). We define two queries Q_{loss} and Q_{gain} to search for loss and gain events, respectively. The two queries are defined as follows:

$$\begin{aligned} Q_{\text{loss}} &= ((\geq, 0.20, =, \text{loss}),\ (\leq, 0.10, =, \text{gain})), \\ Q_{\text{gain}} &= ((\geq, 0.20, =, \text{gain}),\ (\leq, 0.10, =, \text{loss})). \end{aligned}$$

These two queries specify that we are searching for regions with at least 20% of cases (HPV-positive) and at most 10% of controls (HPV-negative) having a gain or loss, respectively. Merging of adjacent regions was enabled with a distance threshold of 0 base pairs again. The Q_{loss} query found 23 regions in chromosomes 2, 3, 4, 5, and 17 (Table 2). The q-values ranged from 0.0120 to 0.911.

Table 1 CNVRs extracted from penile cancer data set using the statistical model with *p*-value ≤ 0.05

Chr	Start	End	Type	*P*-value	Q-value	HPV-pos. freq. (%)	HPV-neg. freq. (%)
4	9729740	24650257	Loss	0.0341	0.0955	21.4	0-3.7
2	230554659	234415376	Loss	0.0387	0.3650	28.6-35.7	0
17	15537019	18617236	Loss	0.0341	0.4632	21.4	0-3.7
2	204245506	207036312	Loss	0.0341	0.5235	21.4	0-3.7
2	237307835	238724893	Loss	0.0387	0.6505	28.6-35.7	0-3.7
9	33911175	34589574	Gain	0.0387	0.6977	28.6	3.7
4	40058630	40957235	Loss	0.0341	0.6977	21.4	0
17	19272468	20059509	Loss	0.0341	0.7198	21.4	0
17	7496965	8209436	Loss	0.0341	0.7295	21.4	3.7
8	37606006	38160563	Gain	0.0341	0.7325	21.4	0
19	51889824	52236621	Gain	0.0341	0.7465	21.4	0
16	779112	798699	Gain	0.0387	0.7907	28.6	3.7
5	130994540	131251586	Loss	0.0341	0.8083	21.4	0
2	218474389	218676793	Loss	0.0341	0.8163	21.4	0
3	53066401	53145339	Loss	0.0350	0.8387	35.7	0
2	240601010	240608450	Loss	0.0387	0.8415	28.6	3.7

Regions are sorted by q-value. The two rightmost columns contain the frequency of variations of the type corresponding to the type of event (column 4) for the case and control groups, respectively

The only region with a q-value < 0.05 was a large region in chromosome 2 in which the frequency of copy number loss was between 21.4 and 35.7% for the HPV-positive samples and between 0 and 7.4% for the HPV-negative samples. The Q_{gain} query found 15 regions in chromosomes 6, 7, 8, 9, 16, 17, 19 and 20 (Table 3). The q-values ranged from 0.276 to 0.871, and thus none of the regions had a significant q-value.

In some cases, one can tighten the thresholds in the query in order to achieve more precise results. If we use the query $Q^*_{\text{loss}} = ((\geq, 0.30, =, \text{loss}), (\leq, 0.05, \neq, \text{normal}))$, we instead find two smaller regions in chromosome 2 (Table 4) with q-values 0.038 and 0.047. Both regions a part of the large loss event found with Q_{loss}, but with a stronger association to HPV-positive status.

Discussion

We used CoNVaQ to search for genomic regions where the occurrence of copy number variations was significantly associated to HPV status. The statistical model found 16 CNVRs with p-value < 0.05. However, none of the regions were reported to have a significant q-value (< 0.05) after permutation testing. This means that for all of the found CNVRs, if the samples are randomly assigned to the two groups, we will likely see an equally large region of same type with p-value < 0.05. This could suggest that for these regions, further validation is necessary to determine whether they are in fact associated with the phenotype. The query-based model found 23 regions for Q_{loss} and 15 regions for Q_{gain}. One region had a significant q-value, namely region in chromosome 2 with copy number loss associated with positive HPV status.

The large discrepancy between the p-values and q-values reported by the statistical model suggests that looking at the individual regions is not sufficient to determine whether a CNVR is indeed significantly associated to a phenotype. The cohort used in our analysis consists of only 41 samples which is evidently too few to determine significance with high confidence. It

Table 2 CNVRs extracted from penile cancer data set using the query-based model with the Q_{loss} query

				HPV-pos.	HPV-neg.
Chr	Start	End	Q-value	loss (%)	loss (%)
2	204245506	240688770	0.0120	21.4-35.7	0-7.41
4	9729740	24650257	0.2037	21.4	0
4	40058630	47037351	0.3377	21.4	0-7.41
5	140934280	147709268	0.3785	21.4-28.6	3.7-7.41
5	60380235	66930240	0.3835	21.4	3.7-7.41
5	76699391	82660330	0.3955	21.4	7.41
17	13330532	18617236	0.4525	21.4	0-7.41
5	126804776	131251586	0.4948	21.4	0-7.41
5	72910129	75996874	0.5863	21.4	7.41
17	19272468	22200000	0.6252	21.4	0-7.41
17	7262327	9931292	0.6633	21.4	0-7.41
17	567713	3136246	0.6793	21.4	7.41
3	46788991	49028973	0.6957	28.6	7.41
17	4566909	6490288	0.7335	21.4	3.7-7.41
5	108507593	110029337	0.7552	21.4-28.6	7.41
2	241709000	242951149	0.7762	21.4	3.7-7.41
3	52397990	53145339	0.8732	21.4-35.7	7.41
5	148184554	148894433	0.8770	21.4	3.7
3	14191317	14755952	0.8862	21.4	7.41
5	140011232	140459066	0.8912	21.4	3.7-7.41
5	92966197	93334626	0.8935	21.4	7.41
5	55167345	55283138	0.9103	21.4	7.41
5	122689458	122739532	0.9113	28.6	7.41

Searching for regions with a loss of copy number in at least 20% of cases and at most 10% of controls. The two rightmost columns contain the frequency of copy nuber loss for the case and control groups, respectively

Table 3 CNVRs extracted from penile cancer data set using the query-based model with the Q_{gain} query

				HPV-pos.	HPV-neg.
Chr	Start	End	Q-value	gain (%)	gain (%)
9	70091642	90376569	0.2762	21.4	7.41
9	13045258	21368309	0.4562	21.4	3.7-7.41
9	119273838	124799603	0.4885	21.4-28.6	7.41
9	33871385	37132743	0.5633	21.4-28.6	3.7-7.41
20	43179131	45592378	0.6395	21.4	7.41
6	13768413	16036680	0.6512	21.4	3.7
19	50012054	52236621	0.6655	21.4	0-7.41
9	91369725	93273440	0.7047	28.6	7.41
16	463336	1664507	0.7648	21.4-28.6	3.7-7.41
17	77682387	78774742	0.7680	21.4	7.41
6	31191394	31875972	0.7963	21.4	3.7
8	37606006	38160563	0.8532	21.4	0
20	33204027	33688992	0.8565	21.4	7.41
7	54865060	55296001	0.8600	21.4	7.41
20	42433913	42557332	0.8710	21.4	7.41

Searching for regions with at least 20% of cases having a gain and at most 10% of controls having any kind of variation. The two rightmost columns contain the frequency of copy number gain for the case and control groups, respectively

illustrates the need for doing proper permutation testing and reporting q-values along with the standard *p*-values. One factor, however, is that the q-value statistic is computed over the entire genome. If the analysis is restricted to a single chromosome, the q-values generally become much smaller. However, given that association studies generally aim to find any variation in any chromosome associated to a trait or phenotype, we believe the q-value statistic should be computed over the entire genome.

Our method currently supports only discrete labels for CNV calls. When the copy number for each CNV is discretized before analysis, information that may potentially be important is discarded. Future versions of CoNVaQ will be extended to also support numerical values for CNV calls in addition to the three categories currently supported (gain, loss and LOH). This would enable determining significance using regression analysis (e.g. linear and logistic regression) and statistical hypothesis tests such as Student's *t*-test or the Mann-Whitney *U* test.

Our platform does not currently support uploading and processing raw genomic data. While this would improve user-friendliness, we believe this is currently out of scope for our method. For now we believe that quality control and CNV calling is best handled by the software tools provided with the CNV detection platforms. By working with processed CNV data instead, it makes our tool agnostic to the detection method used, and can thus be used with both aCGH and next-generation sequencing data.

Conclusions

In this paper we presented CoNVaQ, a web tool for copy number variation-based association studies. CoNVaQ implements two models: a statistical model using Fisher's exact test for significance estimation and a novel query-based model that extract CNV regions matching some user-specified query. Our method provides a secondary significance method by computing an empirical q-value by repeated random permutation of the samples among the two groups. CoNVaQ is provided as a web tool accessible online, making it very simple to use and

Table 4 CNVRs extracted from penile cancer data set using they query-based model with the Q^*_{loss} query

				HPV-positive		HPV-negative	
Chr	Start	End	Q-value	Gain (%)	Loss (%)	Gain (%)	Loss (%)
2	233245552	234392866	0.0382	0	35.7	0	3.7
2	237307836	238138001	0.0473	0-7.14	35.7	0	3.7

Searching for regions with a copy number loss in at least 30% of cases and any kind of variation in at most 5% of controls. Columns 5 and 6 show the frequency of copy number gain and loss for the case group, and columns 7 and 8 shows the frequency of gain and loss for the control group requiring no additional software besides a web browser. Through the web interface we also provide gene set enrichment analysis to easily determine whether a set of CNV regions are associated with GO terms, molecular pathways or diseases. We used CoNVaQ to analyze a data set containing segmented CNV calls for 41 penile cancer patients categorized into HPV positive and HPV negative. While the standard statistical analysis found regions with significant p-value (< 0.05), no region had a significant q-value as well. Q-values were observed to generally be significantly larger than the p-values for the corresponding regions suggesting that looking at each region in isolation is not sufficient for determining significance. While the q-value measure appears to be conservative, we argue that a global significance measure is necessary to reduce type I errors.

Availability and requirements

Project name: CoNVaQ
Project home page: https://convaq.compbio.sdu.dk
Archived version: DOI: `10.5281/zenodo.1217803` (backend), DOI: `10.5281/zenodo.1217898` (frontend).
Operating system(s): Platform independent
Programming language: R and C++
Other requirements: Browser supporting HTML5 and Javascript
License: MIT
Any restrictions to use by non-academics: None

Abbreviations

CNV: Copy number variation; CNVR: Copy number variation region; GWAS: Genome-wide association study; LOH: Lack of heterozygosity; SNP: Single nucleotide polymorphism

Acknowledgements

We acknowledge Ariane Busso-Lopes for providing the aCGH data used in the analysis. We acknowledge the fruitful discussions with Florian Markowetz from Cambridge University.

Funding

SJL and JB are grateful for financial support from JB's VILLUM Young Investigator grant nr. 13154.

Availability of data and materials

CoNVaQ as well as text data and tutorials etc. are available online at https://convaq.compbio.sdu.dk.

Authors' contributions

SJL and JB developed the statistics and algorithms behind CoNVaQ. SJL implemented CoNVaQ. LMC and SRR provided the data sets and helped testing the web service and methodology, and contributed with fruitful biomedical discussions. JB and SRR supervised the whole project. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, DK-5230, Odense, DK, Denmark . [2]Chair of Experimental Bioinformatics, Wissenschaftszentrum Weihenstephan, Technical University of Munich, Maximus-von-Imhof-Forum 3, 85354, Freising-Weihenstephan, DE, Germany . [3]Department of Climcical Genetics, Vejle Hospital and Institute of Regional Health Research, University of Southern Denmark, Beriderbakken 4, DK-7100, Vejle, DK, Denmark . [4]International Center for Research (CIPE), A.C. Camargo Cancer Center, Tagua 440, 01508-010, Sao Paulo, BR, Brazil .

Received: 12 December 2017 Accepted: 26 April 2018
Published online: 18 May 2018

References

1. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, Jorde LB, Posukh OL, Sahakyan H, Watkins WS, Yepiskoposyan L, Abdullah MS, Bravi CM, Capelli C, Hervig T, Wee JTS, Tyler-Smith C, van Driem G, Romero IG, Jha AR, Karachanak-Yankova S, Toncheva D, Comas D, Henn B, Kivisild T, Ruiz-Linares A, Sajantila A, Metspalu E, Parik J, Villems R, Starikovskaya EB, Ayodo G, Beall CM, Rienzo AD, Hammer MF, Khusainova R, Khusnutdinova E, Klitz W, Winkler C, Labuda D, Metspalu M, Tishkoff SA, Dryomov S, Sukernik R, Patterson N, Reich D, Eichler EE. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):3761. https://doi.org/10.1126/science.aab3761.
2. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172–83. https://doi.org/10.1038/nrg3871.
3. Gonzalez E. The influence of CCL3l1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science. 2005;307(5714): 1434–40. https://doi.org/10.1126/science.1101160.
4. Hollox EJ, Huffmeier U, Zeeuwen PLJM, Palla R, Lascorz J, Rodijk-Olthuis D, van de Kerkhof PCM, Traupe H, de Jongh G, den Heijer M, Reis A, Armour JAL, Schalkwijk J. Psoriasis is associated with increased β-defensin genomic copy number. Nat Genet. 2007;40(1):23–5. https://doi.org/10.1038/ng.2007.48.
5. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bölte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BHY, Cochrane L, Corsello C, Crawford EL, Crossett A, Cytrynbaum C, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green A, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Couteur AL, Leventhal BL, Lionel AC, Liu X-Q, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Pilorge M, Piven J, Ponting CP, Posey DJ, Poustka A, Poustka F, Prasad A, Ragoussis J, Renshaw K, Rickaby J, Roberts W, Roeder K, Roge B, Rutter ML, Bierut LJ, Rice JP, Salt J, Sansom K, Sato D, Segurado R, Sequeira AF, Senman L, Shah N, Sheffield VC, Soorya L, Sousa I, Stein O, Sykes N, Stoppioni V, Strawbridge C, Tancredi R, Tansey K, Thiruvahindrapduram B, Thompson AP, Thomson S, Tryfon A, Tsiantis J, Engeland HV, Vincent JB, Volkmar F, Wallace S, Wang K, Wang Z, Wassink TH, Webber C, Weksberg R, Wing K, Wittemeyer K, Wood S, Wu J, Yaspan BL, Zurawiecki D, Zwaigenbaum L, Buxbaum JD, Cantor RM, Cook EH, Coon H, Cuccaro ML, Devlin B, Ennis S, Gallagher L, Geschwind DH, Gill M, Haines JL, Hallmayer J, Miller J, Monaco AP, Jr JIN, Paterson AD, Pericak-Vance MA, Schellenberg GD, Szatmari P, Vicente AM, Vieland VJ, Wijsman EM, Scherer SW, Sutcliffe JS, Betancur C. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466(7304):368–72. https://doi.org/10.1038/nature09146.
6. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW,

Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848–53. https://doi.org/10.1126/science.1136678.

7. Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. Brief Funct Genom. 2015;14(5):352–7. https://doi.org/10.1093/bfgp/elv017.
8. Mitri Z, Constantine T, O'Regan R. The HER2 receptor in breast cancer: Pathophysiology, clinical use, and new advances in therapy. Chemother Res Pract. 2012;2012:1–7. https://doi.org/10.1155/2012/743193.
9. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, MacDonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512–20. https://doi.org/10.1038/nbt.1852.
10. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14(Suppl 11):1. https://doi.org/10.1186/1471-2105-14-s11-s1.
11. Kim J-H, Hu H-J, Yim S-H, Bae JS, Kim S-Y, Chung Y-J. CNVRuler: a copy number variation-based case-control association analysis tool. Bioinformatics. 2012;28(13):1790–2. https://doi.org/10.1093/bioinformatics/bts239.
12. Glessner JT, Li J, Hakonarson H. ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Res. 2013;41(5):64–4. https://doi.org/10.1093/nar/gks1346.
13. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P. Ensembl 2016. Nucleic Acids Res. 2015;44(D1):710–6. https://doi.org/10.1093/nar/gkv1157.
14. Consortium GO, et al. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43(D1):1049–56. https://doi.org/10.1093/nar/gku1179.
15. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45(D1):353–61. https://doi.org/10.1093/nar/gkw1092.
16. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2017;46(D1):649–55. https://doi.org/10.1093/nar/gkx1132.
17. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2014;43(D1):1071–8. https://doi.org/10.1093/nar/gku1011.
18. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2016;45(D1):833–9. https://doi.org/10.1093/nar/gkw943.
19. Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an r/bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2014;31(4):608–9. https://doi.org/10.1093/bioinformatics/btu684.
20. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an r package for comparing biological themes among gene clusters. OMICS: A J Integr Biol. 2012;16(5):284–7. https://doi.org/10.1089/omi.2011.0118.
21. Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol. 2002;64(3):479–98.
22. Busso-Lopes AF, Marchi FA, Kuasne H, Scapulatempo-Neto C, Trindade-Filho JCS, de Jesus CMN, Lopes A, Guimarães GC, Rogatto SR. Genomic profiling of human penile carcinoma predicts worse prognosis and survival. Cancer Prev Res. 2015;8(2):149–56. https://doi.org/10.1158/1940-6207.CAPR-14-0284.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC