CNV-P: A machine-learning framework for predicting high confident copy number variations

Taifu Wang[1*], Jinghua Sun[1,2*], Xiuqing Zhang[1,2,3], Wen-Jing Wang[1#], Qing Zhou[1#]

[1] BGI-Shenzhen, Shenzhen 518083, China

[2] College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049,China

[3] Guangdong Enterprise Key Laboratory of Human Disease Genomics, Beishan Industrial Zone, Shenzhen, 518083, China

Corresponding Author:

Qing Zhou [1]

China National GenBank, Jinsha Road, Shenzhen, Guangdong, 518120, China

Email address: zhouqing1@genomics.cn

Abstract

Background: Copy-number variants (CNVs) have been recognized as one of the major causes of genetic disorders. Reliable detection of CNVs from genome sequencing data has been a strong demand for disease research. However, current software for detecting CNVs has high false-positive rates, which needs further improvement.

Methods: Here, we proposed a novel and post–processing approach for CNVs prediction (CNV-P), a machine-learning framework that could efficiently remove false-positive fragments from results of CNVs detecting tools. A series of CNVs signals such as read depth (RD), split reads (SR) and read pair (RP) around the putative CNV fragments were defined as features to train a classifier.

Results: The prediction results on several real biological datasets showed that our models could accurately classify the CNVs at over 90% precision rate and 85% recall rate, which greatly improves the performance of state-of-the-art algorithms. Furthermore, our results indicate that CNV-P is robust to different sizes of CNVs, as well as the platforms of sequencing.

Conclusions: Our framework for classifying high-confident CNVs could improve both basic research and clinical diagnosis of genetic diseases.

Introduction

Copy number variations (CNVs) are one of the genetic variations and sequence polymorphisms that widely exist in the human genome. Research shows that CNVs are closely related to the pathogenesis and development of many human diseases such as autism, Parkinson and other neurological diseases (Hollox et al. 2008; Pankratz et al. 2011; Rosenfeld et al. 2010; Sebat et al. 2007). Therefore, accurate detection of CNVs is essential for the diagnosis and research of such diseases.

With the rapid development of high-throughput sequencing technology, genomic sequencing-based technology for CNVs detection has gradually become a leading method owing to its high speed, high resolution, and high repeatability. Many sequencing-based CNVs detection methods have been proposed (Kosugi et al. 2019; Pirooznia et al. 2015; Zhao et al. 2013). Typical CNVs detection approaches mainly utilize three signatures to detect CNVs: read depth (RD), read pairs (RP), split read (SR) (Pirooznia et al. 2015). RD means the number of reads that encompass or overlap CNVs. For example, a deletion indicates a decrease in the average depth of this area. RP refers to the distribution of the insert size of the sequenced library. If the mapping distance of read pairs significantly deviates from the average value of the sequencing library, such discordant alignment features herald the occurrence of CNVs. SR

49 indicates the split (soft-clipped) alignment features of reads that span CNVs. The initial
50 strategies for detecting CNVs mainly focused on one of these features (Abyzov et al. 2011; Chen
51 et al. 2009; Medvedev et al. 2010). Most of the approaches have high false-positive rates because
52 of the noises of sequencing data, such as sequencing error and artificial chimeric reads.
53 Ambiguous mapping of reads from repeat- or duplication-rich regions also decreases the
54 accuracy of CNVs (Kosugi et al. 2019; Teo et al. 2012). Consequently, tools integrating multiple
55 features to detect CNVs have been gradually developed (Bartenhagen & Dugas 2016; Layer et
56 al. 2014; Rausch et al. 2012), while their performance still needs be further modified (Kosugi et
57 al. 2019).
58   To identify high confident CNVs, a commonly used strategy is setting a cutoff value or
59 applying various statistical distributions to filter fragments. This strategy greatly depends on the
60 expertise of researchers and their subjective assumptions about the analyzed data. Another
61 strategy uses the intersection of CNVs generated by two or more algorithms. However, due to
62 various CNV-property-dependent and library-property-dependent features used by different
63 detection methods, they usually provide inconsonant results. Thus, a large number of potentially
64 true CNVs could be discarded. Additionally, some tools, such as MetaSV(Mohiyuddin et al.
65 2015), Parliament2 (Zarate et al. 2020) and FusorSV (Becker et al. 2018), use the method of
66 integrating and merging CNVs from multiple software. These approaches require output results
67 of several certain tools, usually more than four software, while reanalyzing CNVs using their
68 default methods is impractical and time-consuming. Recently, some software using machine
69 learning model to detect or predict CNVs have been developed, such as GATK-SV
70 (Werling et al. 2018) and CNV- JACG (Zhuang et al. 2020), it is still necessary to develop more
71 accurate tools.
72   Here, we developed a machine-learning framework for CNVs prediction (CNV-P), aiming
73 to accurately predict CNVs from the results of present software. CNV-P collected three
74 aforementioned signatures (RD, RP and SR) and other information of the putative CNVs. The
75 results of our model on real data demonstrate that CNV-P greatly improves the performance of
76 state-of-the-art algorithms

77

Materials & Methods

Data download and preprocessing

80   The gold-standard sets of CNVs from 9 individuals (NA19238, NA19239, NA19240,
81 HG00512, HG00513, HG00514, HG00731, HG00732, HG00733) were downloaded from
82 *Chaisson et al* (Chaisson et al. 2019). The whole genome sequencing (WGS) data (~30x) of

83   these 9 individuals were downloaded from the National Center for Biotechnology Information

84   (NCBI) with an accession number SRP159517 (***Table S1, S2***). For external validation samples,

85   the sequencing data of NA12878 and HG002 were also downloaded from NCBI with accession

86   numbers SRP159517 and SRP047086 respectively. The gold-standard CNVs of NA12878 were

87   generated by three data sets: the Database of Genomic Variants

88   (http://dgv.tcag.ca/dgv/app/home?ref=GRCh37/hg19) (R. et al. 2013), the 1000 Genomes Project

89   phaseIII (https://ftp.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map) (Sudmant et al.

90   2015), and the CNVs of PacBio data from *Pendleton, M. et al* (Pendleton et al. 2015). The gold-

91   standard CNVs of HG002 were downloaded from *Zook, J.M., et al* (Zook et al. 2019).

92   For the above gold-standard sets, we excluded other types of CNVs except for deletion and

93   duplication, removed CNVs shorter than 100bp and merged fragments with over 80% reciprocal

94   overlaps. On average, each sample had more than 10,000 CNVs after processing (***Table S1***). For

95   WGS data, the clean reads after removing adapter and filtering low-quality reads were aligned to

96   the human genome reference (hg19) with bwa (Li 2013) 'mem' command to generate the BAM

97   file. All of these datasets were generated by standard WGS protocol, with libraries of

98   approximate 400bp insert size and average ~30X coverage (***Table S2***).

99

Generate simulated dataset

101   We generate random CNVs (range from 100bp to 100kb) based on a copy of human

102   genome (hg19) using mason2 (Holtgrewe 2010). To avoid the same or similar CNVs between

103   training data and test data, we selected fragments on chromosome 1 and chromosome 2 as

104   training samples and CNVs on chromosome 3 and chromosome 4 as testing samples (more

105   details in ***Table S1***). Then, the paired-end sequencing reads (100bp) from the altered genome

106   was simulated by wgsim (Li 2011), with an insert size of 500bp and 0.001 base error rate.

107

Training set and test set

109   We chose five common software to obtain the initial sets of CNVs for simulated data and

110   the downloaded sequencing data (deletions and duplications): Lumpy (Layer et al. 2014), Manta

111   (Chen et al. 2015), Pindel (Ye et al. 2009), Delly (Rausch et al. 2012) and breakdancer (Chen et

112   al. 2009). The details of running parameters were shown in the supplemental methods section.

113   The original CNVs were then performed as follows: 1. Removed other types of CNVs except for

114   deletion and duplication. 2. Removed CNVs with > 10 bp overlapped with N region of human

115   genome (download from http://genome.ucsc.edu/). 3. Merged CNVs with ≥80% reciprocal

116   overlaps and kept the union part of fragments. 4. Removed CNVs that less than 100bp. Then, we

117   labeled these treated CNVs as either 'True' or 'False' based on their overlapped part with gold-
118   standard CNVs. CNVs having ≥80% reciprocal overlap with the gold-standard CNVs in
119   simulated data were labeled as "True" and the cutoff was set to ≥50% for sequencing data. We
120   then selected data of 6 individuals as a training set and the other 3 samples as a test set, including
121   two dependent validation datasets (more details in *Table S1, S2*).
122

123   **Feature extraction**

124   We chose commonly used signals by detection tools as features in our training model, such
125   as read depth, information of paired and spited read, mapping quality and GC content of CNVs,
126   as well as all these features around CNV's boundaries (*Table S3*). Training features were
127   obtained from the alignment results (*Fig. 1*) in BAM format, which was generated by a read
128   aligner that supports partial read alignments, such as BWA-MEM (Li 2013). For read depth-
129   based and GC content-based features, we computed the read depth and GC rate of three regions:
130   500 bp upstream and downstream of the left breakpoint L_{b1k}, 500 bp upstream and downstream
131   of the right breakpoint R_{b1k}, and the region from start to end $C_{start-end}$. Read depth was calculated
132   by total number of aligned bases divided by the length of the region. We then normalized the
133   read depth by the average coverage of entire genome and processed log2 transformation to
134   eliminate the impact of fluctuations in sequencing depth. GC content was also calculated in these
135   three regions. Thus, using read depth and GC content of the three local regions (L_{b1k}, R_{b1k} and
136    $C_{start-end}$), six features were defined. Split-read, read pair and mapping quality were computed for
137   two regions: L_{b1k} and R_{b1k}. Split read-based features were defined as the number of clipped
138   reads within the area L_{b1k} or R_{b1k}. Read pair–based features were defined as the number of outlier
139   reads pair within L_{b1k} or R_{b1k}. Normally, The insert size of a normal paired-end read should be
140   within $m_{is} \pm n\sigma_{is}$, where m_{is} and σ_{is} are the median and standard deviation of insert size,
141   respectively, and n is the number of standard deviation from the median(we set is to 3). In
142   addition to aberrant insert size, we also calculated the number of reads without pair within the
143   area L_{b1k} or R_{b1k}. The features of mapping quality were defined as the number of reads with
144   mapping quality <10 within L_{b1k} or R_{b1k}. Finally, we also normalized the value of split reads, read
145   pair and mapping quality according to the mean value of genome coverage. Besides, we included
146   the size and type of CNVs as training features, since the efficacy of CNVs could vary for
147   different size ranges and types (duplication/deletion).
148

149   **Comparison with CNV-JACG, MetaSV and hard cutoff method**

150   We compared the performance of CNV-P with that of CNV- JACG (Zhuang et al. 2020),
151   MetaSV (Mohiyuddin et al. 2015) and hard cutoff method in the same datasets. Since MetaSV
152   currently does not support Delly's output, only four CNV detection tools (Lumpy, Manta, Pindel,
153   and breakdancer) were taken into consideration. CNV-JACG was conducted running with default
154   parameters (details in supplementary methods). MetaSV was carried out with complete mode.
155   For hard cutoff method, we used SR and RP as the evidence to support the existence of CNVs,
156   therefore, the number of SR and RP greater than 2, 5, and 10 were set as hard cutoff to evaluate.
157   SURVIVOR(Jeffares et al. 2017) was used to merge fragments with 80% overlap after filtering
158   by CNV-P, CNV- JACG, MetaSV and hard cutoff method.

159

160   Methodology evaluation

161   we calculated the classifier performance on the test dataset in terms of precision and recall
162   (TP: true positive, TN: true negative, FP: false positive, FN: False negative)

163    $\text{Precision} = \dfrac{\text{TP}}{\text{TP} + \text{FP}}$

164    $\text{Recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}}$

165    $\text{F1 score} = \dfrac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

166   Also, we plotted the ROC with the AUC for model evaluation. ROC curves were drawn
167   based on a series of false positive rates (FPR) and true positive rates (TPR).

168

169   Results

170   Study overview

171   In this study, we built a random forest (RF) framework for the CNVs prediction base on
172   both simulated and real datasets (*Fig. 1A*). Firstly, we identified CNVs using five common tools
173   (Lumpy, Manta, Pindel , Delly and breakdancer). For each set, we removed CNVs with low
174   quality or locating on the N region of the human genome (details in **Methods**). Secondly, we
175   labeled CNVs as either "True" or "False" based on a 50% reciprocal overlap with the gold-
176   standard CNVs in real data and 80% reciprocal overlap in simulated data respectively (details in
177   **Methods**). Next, we extracted the signatures around these CNVs such as RD, SR and RP as
178   training features from alignment results (*Fig. 1B*).
179   We then split the data set into a training set and a test set. Based on the training set, we
180   trained a RF classifier to identify CNVs as "true" or "false". We performed 10-repeated 10-fold

181 cross-validation for optimal parameter selection and used the receiver operating characteristic
182 (ROC) curve to quantify the prediction performance. Next, we evaluated the robustness of our
183 models on test data from multiple aspects, such as sizes of CNVs and platforms of raw
184 sequencing data. We also compared the performance of Support Vector Machine (SVM) and
185 Gradient Boosting classifier (GBC) with our random forest model. Finally, we validated our
186 model on two extra data sets.

187

188 **Performance of CNV-P on a simulated dataset**

189 We train RF, GBC, and SVM classifiers for CNV prediction based on a simulated dataset
190 (details in **Methods**). The results show that there was a significant improvement after CNV-P
191 prediction compared with the original CNV results. The precision of CNVs produce by each
192 CNV detection tools improved from 49.58% to 99% with almost zero loss recall rate (***Fig. 2A,***
193 ***B***). Compared with GBC and SVM classifiers, RF was slightly superior. The RF classifiers for 5
194 tools achieved comparable performance since an average increasing of F1-score was about
195 14.55% for Lumpy, 14.34% for Manta, 13.84% for Pindel, 11.10% for Delly and 16.16% for
196 breakdancer (***Fig. 2C***).

197

198 **Performance of CNV-P on a real dataset**

199 In this part, we trained RF classifier for the five selected tools respectively based on real
200 samples. The 10-repeated 10-fold cross-validation was performed for optimal parameter
201 selection (***Fig. S1***). The overall diagnostic ability of each classifier was measured as the area
202 under the receiver operating characteristic curve (AUC) for the test dataset. The highest value of
203 AUC was 97.10% for the model of Lumpy while the model for Pindel had the smallest value of
204 93.62% (***Fig. 3A***). Each classifier accurately classified the CNVs as either true or false at 91.76-
205 95.17% precision and 87.75-96.54% recall rate (***Fig. 3B***). After processing by CNV-P, a large
206 number of false-positive CNVs were removed, and the majority of true CNVs were remained
207 (***Fig. 3C***).
208 To dissect the principle of the CNV-P classifier, we assessed the relative importance of each
209 feature for corresponding classifiers. As expected, for all classifiers, read-depth provided the
210 most discriminatory power to make accurate predictions (***Fig. S2***). However, the second
211 important feature was inconsistent between different classifiers. It was probably due to various
212 detection algorism these tools used.
213 To evaluate the robustness of CNV-P, we trained each model on various proportions of
214 training data (from 10% to 90% in increments of 20%). The results showed a steady

215 improvement in accuracy (precision and recall rate) with an increase in the number of training
216 data (**Fig. S3**).
217   We further assessed the performance of CNV-P for CNVs of different sizes. We divided
218 CNVs into three sets based on their length: CNV_S (100 bp to 1 kb; bp: base pair, kb: kilobase),
219 CNV_M (1 kb to 100 kb) and CNV_L (>100kb). The overall precisions were greatly improved,
220 comparing with the raw CNVs achieved by the corresponding software (**Fig. 3D**). We noticed
221 that almost all precision and recall rates of CNV_S and CNV_M were over 90%, while theses
222 values of CNV_L were slightly lower. These results are probably caused by the insufficient
223 number of CNV_L in our training data.
224   We also profiled the distribution of predicted probability scores for all CNVs within a
225 different size range. Since CNVs with a probability score >0.5 were classified as true in our
226 CNV-P prediction results, we found that the threshold of 0.5 distinguished true and false CNVs
227 very well (**Fig. S4**). Besides, the probability scores could be used as a measurement of
228 confidence for a certain fragment of CNVs, which would provide support evidence in further
229 analysis.
230   Furthermore, we implanted two additional models, GBC and SVM, to train CNV-P
231 classifiers. Comparing the precision and recall values, as well as the result of ROC curve, we
232 found they had comparable performance (**Fig. S5**). Still, the RF classifier was recommended as
233 the first choice with a slight superiority.
234

Prediction on external data sets

236   To further evaluate the performance of CNV-P, we implemented our models on two
237 independent WGS datasets of NA12878 and HG002 (***Table S1***). Since we had proved that
238 increasing the size of training data could improve the accuracy of our model (***Fig. S3***), the final
239 classifiers were trained on both the training set and test set mentioned above. Consistent with the
240 above results, CNV-P produced the optimal performance with AUCs of 0.89-0.95 in NA12878
241 (***Fig. 4A***). Most of the false-positive CNVs were removed with a loss of a small number of true
242 positive fragments (***Fig 4B, C***). Likewise, our approach had a similar performance on sample
243 HG002 (***Fig. 4D-F***).
244   We next compared CNV-P with other post-process tools for CNV filtering, including CNV-
245 JACG and MetaSV. We also included commonly used hard filtering method, setting cut-off of
246 SR and RP number for each CNV. We applied various filtering approaches on NA12878 and
247 HG002, and then evaluated fragments using gold-standard CNVs of these two samples. Our
248 results showed that CNV-P had the highest F1-score among all the post-process methods (***Table***
249 ***1***).

250   Besides, we evaluated the performance of our approach in data generated from multiple
251   sequencing platforms. With precision of 91.6-96.8% and recall rates of 84.1-94% (***Fig. 5***), CNV-
252   P showed similar performance on sequencing data generated by BGI-500. Moreover, in addition
253   to the trained classifiers for the above five software, we provided extra modules in our approach
254   for training and predicting if CNVs were detected by other tools. These results suggest that our
255   approach is suitable for CNVs generated from multiple sequencing platforms and detecting
256   software.
257

Discussion

259   Detecting CNVs from WGS is error-prone because of short-length reads and library-
260   property-dependent bias [5]. Inflated false positive makes it a big challenge for researchers to
261   identify clinically relevant CNVs, as it is time- and money-consuming to validate a large amount
262   of false positive CNVs. To solve this problem, we develop CNV-P, an effective machine-
263   learning-based framework to acquire high-confident CNVs. Instead of handling the shortcomings
264   of existing methods by developing another detecting algorithm, CNV-P focuses on providing a
265   reliable set of CNVs from existing detection software. We demonstrate that CNV-P can identify
266   a set of high-confidence CNVs with high precision and recall rates. Moreover, CNV-P is robust
267   to the proportion of variants in training sets, size of CNVs and sequencing platforms, indicating
268   the utility of CNV-P in a variety of clinical or research contexts.
269   Comparing with the conventional method of using hard cutoff, such as a minimum number
270   of supporting reads, to filtering CNV results, CNV-P greatly reduces errors caused by lack of
271   expertise and subjective assumptions. Instead of running default multiple software in advance,
272   CNV-P can make accurate predictions for each tool dependently. In addition to the five
273   commonly used software that we have trained prediction models, we provide an extra module in
274   CNV-P including the function of model training and predicting if CNVs are detected by other
275   tools.
276   However, our models may have weaker power for large-size CNVs, because there are only
277   a small number of large fragments in our training data. Besides of data from healthy individuals,
278   we believe that great improvement could be made to identify large-size true CNVs in the future
279   when more datasets are accumulated.
280

Conclusions

282   CNV-P is a well-performed machine-learning framework for accurately filtering CNVs.
283   CNV-P framework can be applied on CNVs from various detection methods and sequencing

284 platforms, making our framework easy to adopt and customize. CNV-P greatly helps to generate

285 a set of high-confident CNVs, benefiting both basic research and clinical diagnosis of genetic

286 diseases.

287

288 **Availability of data and materials**

289 All data generated or analyzed during this study are included in this published article and its

290 supplementary information files.

291 CNV-P is available at https://github.com/wonderful1/CNV-P.

292

293 **Acknowledgements**

294 The authors thank Dr. Jian Guo for constructive comments on this project and Chen Ye for data

295 download and management.

296

297 **References**

298 Abyzov A, Urban AE, Snyder M, and Gerstein M. 2011. CNVnator: an approach to discover,

299 genotype, and characterize typical and atypical CNVs from family and population genome

300 sequencing. *Genome Res* 21:974-984. 10.1101/gr.114876.110

301 Bartenhagen C, and Dugas M. 2016. Robust and exact structural variation detection with paired-

302 end and soft-clipped alignments: SoftSV compared with eight algorithms. *Brief Bioinform* 17:51-

303 62. 10.1093/bib/bbv028

304 Becker T, Lee WP, Leone J, Zhu Q, Zhang C, Liu S, Sargent J, Shanker K, Mil-Homens A,

305 Cerveira E, Ryan M, Cha J, Navarro FCP, Galeev T, Gerstein M, Mills RE, Shin DG, Lee C, and

306 Malhotra A. 2018. FusorSV: an algorithm for optimally combining data from multiple structural

307 variation detection methods. *Genome Biol* 19:38. 10.1186/s13059-018-1404-6

308 Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ,

309 Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN,

310 Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D, Anantharaman T, Audano

311 PA, Brand H, Cantsilieris S, Cao H, Cerveira E, Chen C, Chen X, Chin CS, Chong Z, Chuang

312 NT, Lambert CC, Church DM, Clarke L, Farrell A, Flores J, Galeev T, Gorkin DU, Gujral M,

313 Guryev V, Heaton WH, Korlach J, Kumar S, Kwon JY, Lam ET, Lee JE, Lee J, Lee WP, Lee

314 SP, Li S, Marks P, Viaud-Martinez K, Meiers S, Munson KM, Navarro FCP, Nelson BJ, Nodzak

315 C, Noor A, Kyriazopoulou-Panagiotopoulou S, Pang AWC, Qiu Y, Rosanio G, Ryan M, Stutz A,

316 Spierings DCJ, Ward A, Welch AE, Xiao M, Xu W, Zhang C, Zhu Q, Zheng-Bradley X, Lowy

317 E, Yakneen S, McCarroll S, Jun G, Ding L, Koh CL, Ren B, Flicek P, Chen K, Gerstein MB,

318 Kwok PY, Lansdorp PM, Marth GT, Sebat J, Shi X, Bashir A, Ye K, Devine SE, Talkowski ME,

319 Mills RE, Marschall T, Korbel JO, Eichler EE, and Lee C. 2019. Multi-platform discovery of

320 haplotype-resolved structural variation in human genomes. *Nat Commun* 10:1784.

321 10.1038/s41467-018-08148-z

322 Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC,

323 Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, and Mardis ER. 2009.

324 BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat*

325 *Methods* 6:677-681. 10.1038/nmeth.1363

326 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, and Saunders CT. 2015. Manta: Rapid detection

327 of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*

328 32:1220-1222.

329 Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, van de Kerkhof PC,

330 Traupe H, de Jongh G, den Heijer M, Reis A, Armour JA, and Schalkwijk J. 2008. Psoriasis is

331 associated with increased beta-defensin genomic copy number. *Nat Genet* 40:23-25.

332 10.1038/ng.2007.48

333 Holtgrewe M. 2010. Mason@   A Read Simulator for Second Generation Sequencing Data.

334 Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bahler J, and

335 Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and

336 reproductive isolation in fission yeast. *Nat Commun* 8:14061. 10.1038/ncomms14061

337 Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, and Kamatani Y. 2019. Comprehensive

338 evaluation of structural variation detection algorithms for whole genome sequencing. *Genome*

339 *Biol* 20:117. 10.1186/s13059-019-1720-5

340 Layer RM, Chiang C, Quinlan AR, and Hall IM. 2014. LUMPY: a probabilistic framework for

341 structural variant discovery. *Genome Biol* 15:R84. 10.1186/gb-2014-15-6-r84

342 Li H. 2011 wgsim - Read simulator for next generation sequencing. *Github Repository [online]*

343 http://githubcom/lh3/wgsim

344 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

345 *arXiv preprint arXiv:13033997*.

346 Medvedev P, Fiume M, Dzamba M, Smith T, and Brudno M. 2010. Detecting copy number

347 variation with mated short reads. *Genome Res* 20:1613-1622. 10.1101/gr.106344.110

348 Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, and Lam HY.

349 2015. MetaSV: an accurate and integrative structural-variant caller for next generation

350 sequencing. *Bioinformatics* 31:2741-2744. 10.1093/bioinformatics/btv204

351 Pankratz N, Dumitriu A, Hetrick KN, Sun M, Latourelle JC, Wilk JB, Halter C, Doheny KF,

352 Gusella JF, Nichols WC, Myers RH, Foroud T, DeStefano AL, Psg P, GenePd Investigators C,

353 and Molecular Genetic L. 2011. Copy number variation in familial Parkinson disease. *Plos One*

354 6:e20988. 10.1371/journal.pone.0020988

355 Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W,

356 Anantharaman T, Hastie A, Dai H, Fritz MH, Cao H, Cohain A, Deikus G, Durrett RE,

357 Blanchard SC, Altman R, Chin CS, Guo Y, Paxinos EE, Korbel JO, Darnell RB, McCombie

358 WR, Kwok PY, Mason CE, Schadt EE, and Bashir A. 2015. Assembly and diploid architecture

359 of an individual human genome via single-molecule technologies. *Nat Methods* 12:780-786.

360 10.1038/nmeth.3454

361 Pirooznia M, Goes FS, and Zandi PP. 2015. Whole-genome CNV analysis: advances in

362 computational approaches. *Front Genet* 6:138. 10.3389/fgene.2015.00138

363 R. MJ, Robert Z, Yuen RKC, Lars F, and Scherer SW. 2013. The Database of Genomic Variants:

364 a curated collection of structural variation in the human genome. *Nucleic Acids Research*:D1.

365 Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, and Korbel JO. 2012. DELLY: structural

366 variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28:i333.

367 Rosenfeld JA, Ballif BC, Torchia BS, Sahoo T, Ravnan JB, Schultz R, Lamb A, Bejjani BA, and

368 Shaffer LG. 2010. Copy number variations associated with autism spectrum disorders contribute

369 to a spectrum of neurodevelopmental disorders. *Genet Med* 12:694-702.

370 10.1097/GIM.0b013e3181f0c5f3

371 Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S,

372 Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K,

373 Lehtimaki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W,

374 Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, and Wigler M. 2007.

375   Strong association of de novo copy number mutations with autism. *Science* 316:445-449.

376   10.1126/science.1138659

377   Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K,

378   Jun G, Fritz MH, Konkel MK, Malhotra A, Stutz AM, Shi X, Casale FP, Chen J, Hormozdiari F,

379   Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton

380   A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal

381   E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S,

382   Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish

383   NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R,

384   Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X,

385   Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Genomes Project C, Mills RE, Gerstein

386   MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, and Korbel JO. 2015. An integrated map

387   of structural variation in 2,504 human genomes. *Nature* 526:75-81. 10.1038/nature15394

388   Teo SM, Pawitan Y, Ku CS, Chia KS, and Salim A. 2012. Statistical challenges associated with

389   detecting copy number variations with next-generation sequencing. *Bioinformatics* 28:2711-

390   2718. 10.1093/bioinformatics/bts535

391   Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM,

392   Markenscoff-Papadimitriou E, Farrell A, Schwartz GB, Wang HZ, Currall BB, Zhao X, Dea J,

393   Duhn C, Erdman CA, Gilson MC, Yadav R, Handsaker RE, Kashin S, Klei L, Mandell JD,

394   Nowakowski TJ, Liu Y, Pochareddy S, Smith

395   L, Walker MF, Waterman MJ, He X, Kriegstein AR, Rubenstein JL, Sestan N, McCarroll SA, N

396   eale BM, Coon H, Willsey AJ, Buxbaum JD, Daly MJ, State MW, Quinlan AR, Marth GT, Roed

397   er K, Devlin B, Talkowski ME, and Sanders SJ. 2018. An analytical framework for whole-

398   genome sequencing association studies and its implications for autism spetrum disorder. *Nat*

399   *Genet* 50:727-736. 10.1038/s41588-018-0107-y

400   Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z. 2009. Pindel: a pattern growth approach to

401   detect break points of large deletions and medium sized insertions from paired-end short reads.

402   *Bioinformatics* 25:2865-2871. 10.1093/bioinformatics/btp394

403   Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, Schatz MC, Boerwinkle

404   E, Gibbs RA, and Sedlazeck FJ. 2020. Parliament2: Accurate structural variant calling at scale.

405   *Gigascience* 9. 10.1093/gigascience/giaa145

406   Zhao M, Wang Q, Wang Q, Jia P, and Zhao Z. 2013. Computational tools for copy number

407   variation (CNV) detection using next-generation sequencing data: features and perspectives. *Bmc*

408   *Bioinformatics* 14 Suppl 11:S1. 10.1186/1471-2105-14-S11-S1

409   Zhuang X, Ye R, So MT, Lam WY, Karim A, Yu M, Ngo ND, Cherny SS, Tam PK, Garcia-

410   Barcelo MM, Tang CS, and Sham PC. 2020. A random forest-based framework for genotyping

411   and accuracy assessment of copy number variations. *NAR Genom Bioinform* 2:lqaa071.

412   10.1093/nargab/lqaa071

413   Zook JM, Hansen NF, Olson ND, Chapman LM, Mullikin JC, Xiao C, Sherry S, Koren S,

414   Phillippy AM, Boutros PC, Sahraeian SME, Huang V, Rouette A, Alexander N, Mason CE,

415   Hajirasouliha I, Ricketts C, Lee J, Tearle R, Fiddes IT, Barrio AM, Wala J, Carroll A, Ghaffari

416   N, Rodriguez OL, Bashir A, Jackman S, Farrell JJ, Wenger AM, Alkan C, Soylev A, Schatz MC,

417   Garg S, Church G, Marschall T, Chen K, Fan X, English AC, Rosenfeld JA, Zhou W, Mills RE,

418   Sage JM, Davis JR, Kaiser MD, Oliver JS, Catalano AP, Chaisson MJ, Spies N, Sedlazeck FJ,

419   and Salit M. 2019. A robust benchmark for germline structural variant detection.

420   *bioRxiv*:664623. 10.1101/664623

Figure 1

The study overview of CNV-P

A) The workflow of CNV-P framework classifyng candidate CNVs as True or False. B) The features we used to train supervised machine learning models

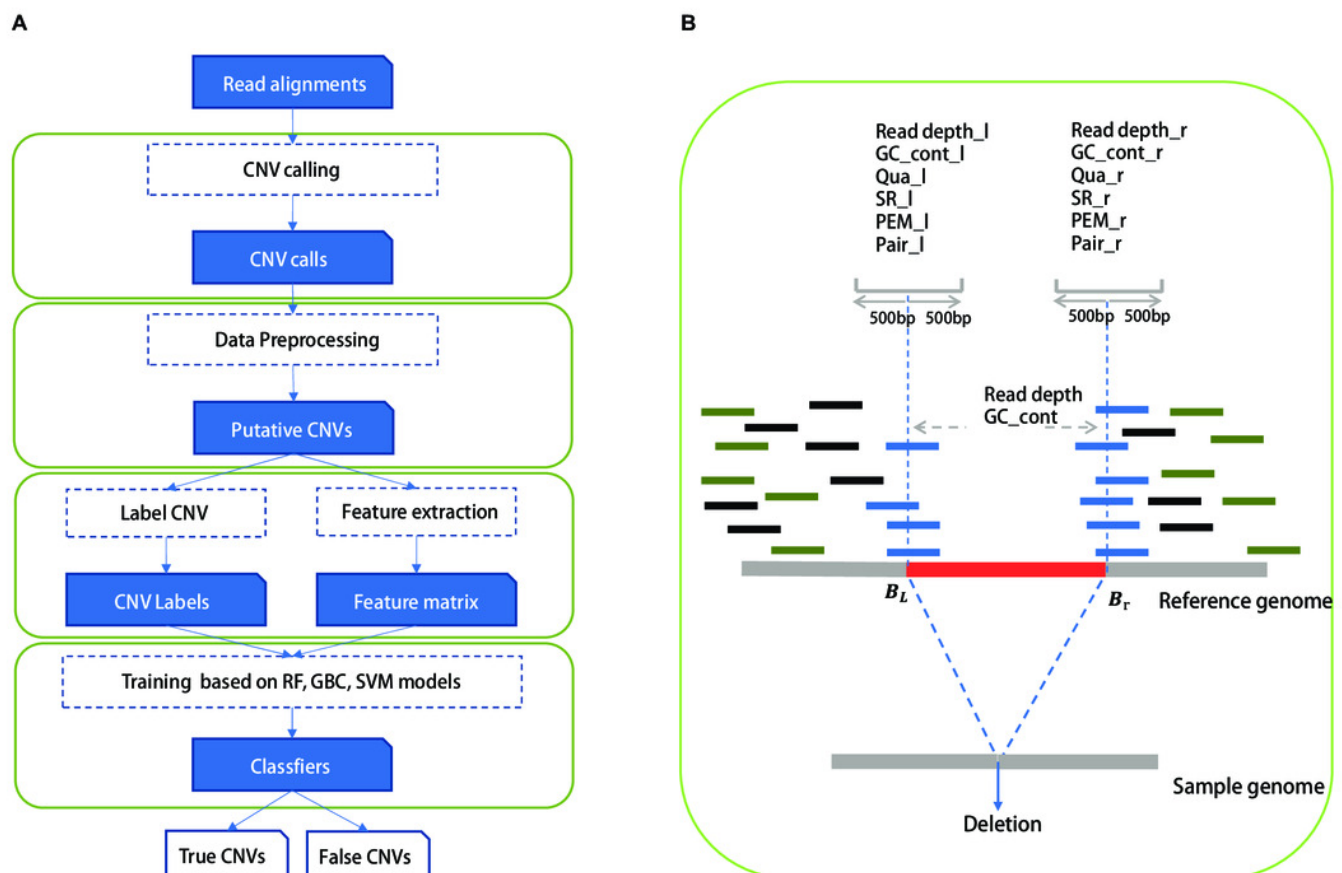



Figure 2

Performance of CNV-P on simulated dataset.

A)The F-score , B) sensitivity, and C) precision over testing simulated dataset.

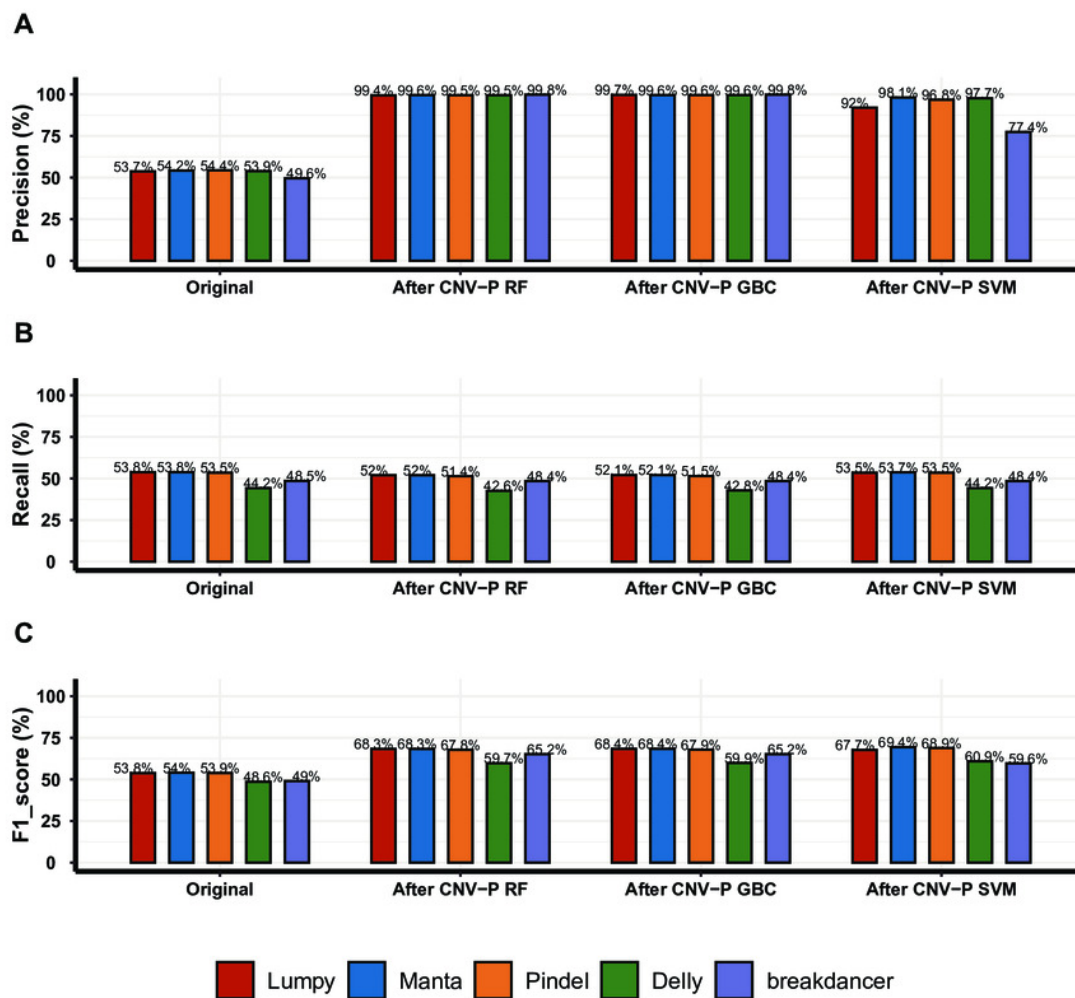



Figure 3

Performance of CNV-P on real dataset.

A) Area Under the receiver operating characteristic Curve (AUC) of CNV-P in 3 test datasets.
B) The precise and recall rate of CNV-P. C) The number of CNVs before and after CNV-P
predicting for five commonly used tools. D) The precise and recall rate of CNV-P at different
size range of CNVs. CNV_S: 100 bp to 1 kb, CNV_M: 1 kb to 100 kb, CNV_L: >100kb.

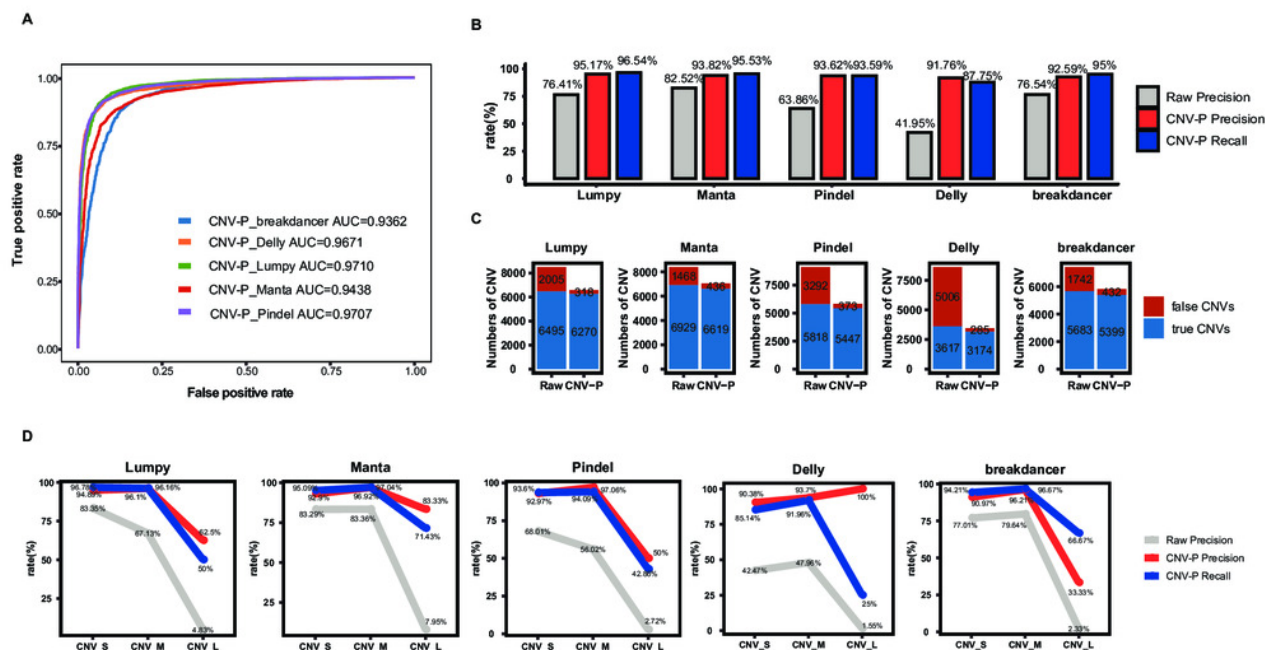



Figure 4

Performance of CNV-P on other validation dataset.

CNV-P detects high-confident CNVs with high precision and recall rates on two independent sequencing datasets from NA12878 (A, B, C) and HG002 (D, E, F). (A, D) Receiver operating characteristic (ROC) curves of CNV-P. (B, E) The precision and recall rate of CNV-P; (C, F) The number of classified CNVs by CNV-P from five commonly used tools.

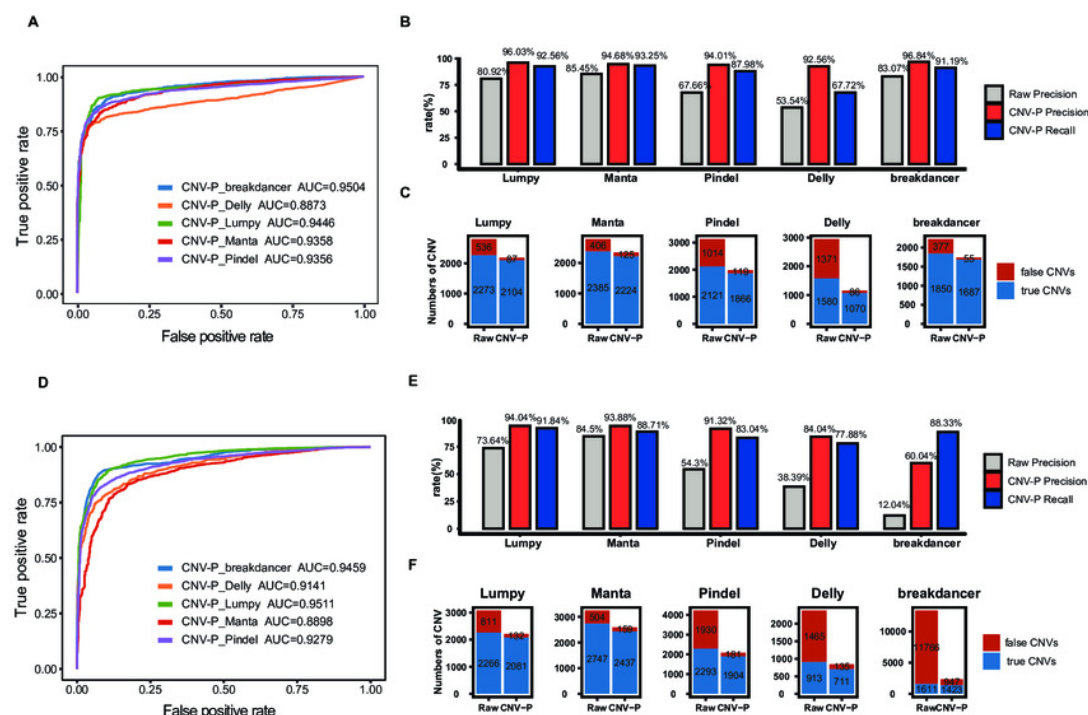
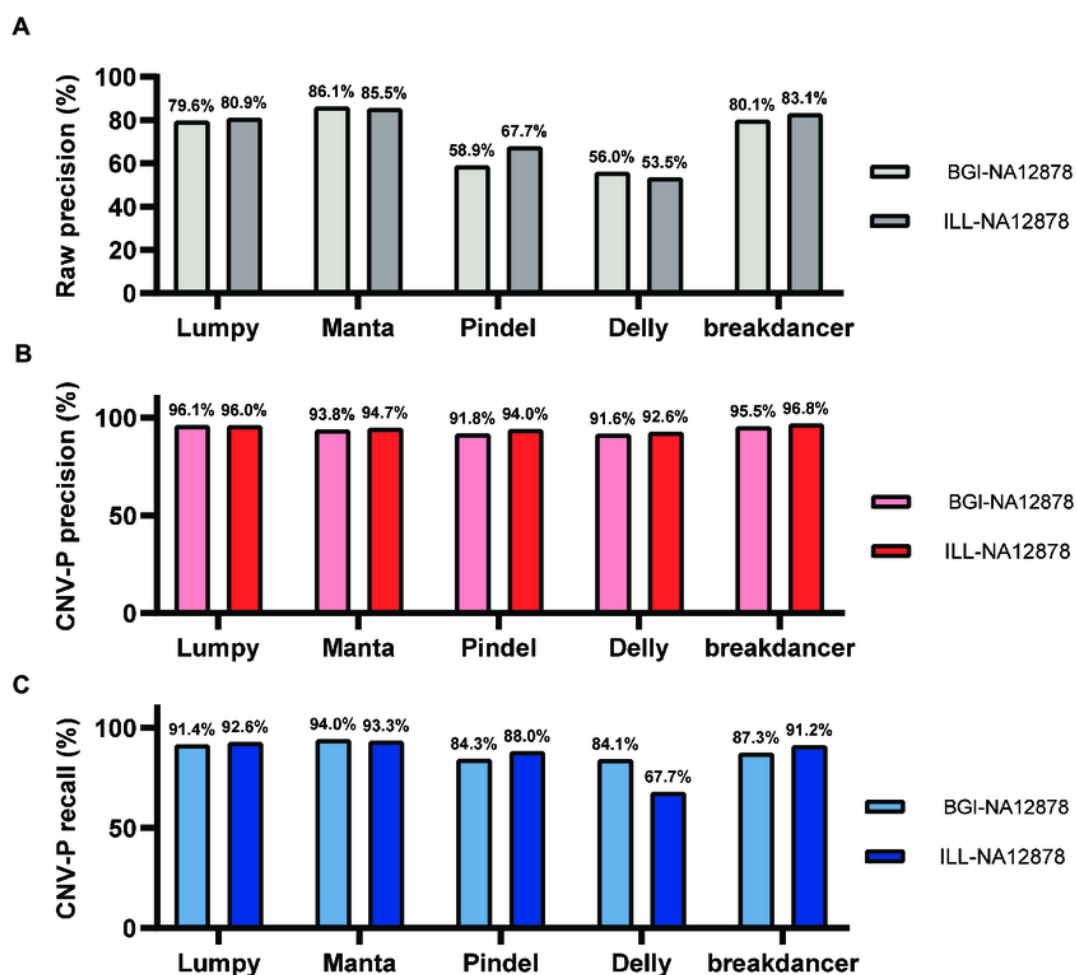



Figure 5

Performance of CNV-P on different sequencing platform.

The precise and recall rate of CNV-P for sample NA12878 using sequencing data generated from BGI-SEQ500 and Illumina. A) The raw precise results; B) Precise rate; C) Recall rate.

Comparison with CNV-JACG, MetaSV and hard cutoff method in NA12878 and HG002

1   Table 1: Comparison with CNV-JACG, MetaSV and hard cutoff method in NA12878 and HG002.

Sample	method	precise	recall	F1-score
NA12878	RAW	0.6032	1.0000	0.7525
	Hard_Cutoff_2	0.6197	0.9792	0.7590
	Hard_Cutoff_5	0.7145	0.8630	0.7818
	Hard_Cutoff_10	0.7780	0.6976	0.7356
	CNV-JACG	0.6828	0.7496	0.7146
	MetaSV	0.7094	0.8817	0.7862
	CNV-P	**0.9007**	**0.7977**	**0.8461**
HG002	RAW	0.2054	1.0000	0.3408
	Hard_Cutoff_2	0.4026	0.9729	0.5695
	Hard_Cutoff_5	0.5740	0.8653	0.6901
	Hard_Cutoff_10	0.6642	0.7482	0.7037
	CNV-JACG	0.5443	0.7076	0.6153
	MetaSV	0.5917	0.8274	0.6900
	CNV-P	**0.7078**	**0.7516**	**0.7290**