Title

SavvyCNV: genome-wide CNV calling from off-target reads

Authors

Thomas W Laver[1], Elisa De Franco[1], Matthew B Johnson[1], Kashyap Patel[1], Sian Ellard[1], Michael N Weedon[1], Sarah E Flanagan[1], Matthew N Wakeling[1]*

1. Institute of Biomedical & Clinical Science, University of Exeter, Exeter, UK

*Corresponding author

Abstract

Current exome and targeted sequencing approaches cannot detect clinically and biologically-relevant copy number variants (CNVs) outside their target area. We present SavvyCNV, a tool which uses off-target read data to call CNVs genome-wide. Up to 70% of sequencing reads from exome and targeted sequencing fall outside the targeted regions - SavvyCNV exploits this 'free data'. Using truth sets generated from genome sequencing data and MLPA, SavvyCNV outperformed four state-of-the-art CNV callers at calling CNVs using off-target reads. We then identified clinically relevant CNVs from a targeted panel using SavvyCNV.

SavvyCNV is freely available from https://github.com/rdemolgen/SavvySuite.

Keywords

Copy number variants, deletions, duplications, next generation sequencing, exomes

Background

Copy number variants (CNVs) are an important class of biological variation. They can cause monogenic disease (1, 2), are associated with polygenic traits (3) and may exert pharmacogenetic effects(4). CNVs are structural rearrangements where bases are gained (duplication) or lost (deletion) from the genome causing an altered copy number compared to the reference.

The importance of CNVs is highlighted by the role they play in many diseases, including cancers (5), autism (6), developmental disorders (7), and heart disease (8). Single or partial gene deletions can cause disease where haploinsufficiency would result in the disease phenotype. For example, both single nucleotide variants and whole gene deletions of *PKD1* cause polycystic kidney disease (1). Duplications can also result in disease as a result of gene disruption at the site of insertion or through increased gene expression. For example, paternal duplication of the chromosome 6q24 region causes neonatal diabetes by overexpression of the imprinted gene *PLAGL1* (2, 9). Larger CNVs are likely to cause syndromic disease as they affect multiple genes. An extreme case is Down syndrome where duplication of chromosome 21 results in characteristic facial features and intellectual disability (10).

CNVs can be detected by a range of methods. In the clinical setting DNA microarrays are routinely used to detect larger rearrangements whilst multiplex ligation-dependent probe amplification (MLPA) is often used to detect single or partial gene CNVs (11). With next generation sequencing (NGS) increasingly employed to investigate genetic variation, the detection of CNVs from NGS data has become increasingly important. While genome sequencing is the optimal method to capture all sequence variation across the genome, due to speed and cost exome sequencing and targeted NGS panels are the most commonly used testing methods, particularly as a first line test in clinical diagnostic laboratories.

Many methods have been published to call CNVs from exome and targeted gene panel data(12). These are designed to detect CNVs within the genes which are targeted by the assay, however biologically interesting and disease causing CNVs will often fall outside of the targeted regions. Existing methods will typically be able to identify if a particular gene is deleted/duplicated however they will not necessarily be able to map the extent of the CNV as the breakpoints will often be located outside the targeted regions.

Current approaches to gene targeting for NGS are imperfect. Samuels *et al* reported that between 40% and 60% of sequence reads generated map outside the target regions (13). This in effect produces ultra-low depth genome sequence data. While there is insufficient information (<1X coverage) to call single nucleotide variants over the untargeted region this 'off target' data can be exploited to call large CNVs by detecting read depth changes over a wide area. The ability to use off-target reads to call CNVs across the genome increases the diagnostic utility of targeted next-generation sequencing panels and also allows for more accurate mapping of breakpoints of CNVs which reside outside of the targeted regions. Previous tools have been designed to detect CNVs in off-target reads from exome data (14) and large targeted panels (15).

We have developed a new tool, SavvyCNV, for calling CNVs from off-target reads. We investigated the utility of SavvyCNV by comparing it to the current state of the art tools for calling CNVs in two different truth sets of CNVs from targeted panels and a truth set from exome data. We then used SavvyCNV in a patient cohort tested with a small targeted gene panel (75 genes) to perform a genome-wide analysis to detect CNVs of clinical relevance.

Results

How much off-target read data is there?

For our small targeted panel (16) of 75 genes, 3.4 (SD 1.6) million reads are sequenced on average per sample. 55% (SD 10%) of these map to off-target regions of the genome. This gives a mean read

depth in off-target regions of 0.065 (SD 0.044). In the exome samples that we used as a benchmarking data set there are an average of 76 (SD 20) million reads per sample with 20.3% (SD 6.6%) off-target, equating to mean read depth of 0.52 (SD 0.20) in off-target regions. This compares to a typical genome sequencing experiment where sufficient reads are sequenced to give >30X mean coverage across the genome.

SavvyCNV can call off-target CNVs from targeted panels

To evaluate SavvyCNV's ability to call off-target CNVs accurately from targeted panel data we benchmarked its performance against a truth set (see Methods) and compared it to four other tools for calling CNVs: GATK gCNV(17), DeCON(18), EXCAVATOR2(14) and CNVkit(15). To prevent bias due to software configuration tuning, we ran all five tools with multiple configurations, and plotted the best results for each tool on a precision-recall graph (Figure 1). The best recall (sensitivity) where precision is at least 50% is shown in Table 1.

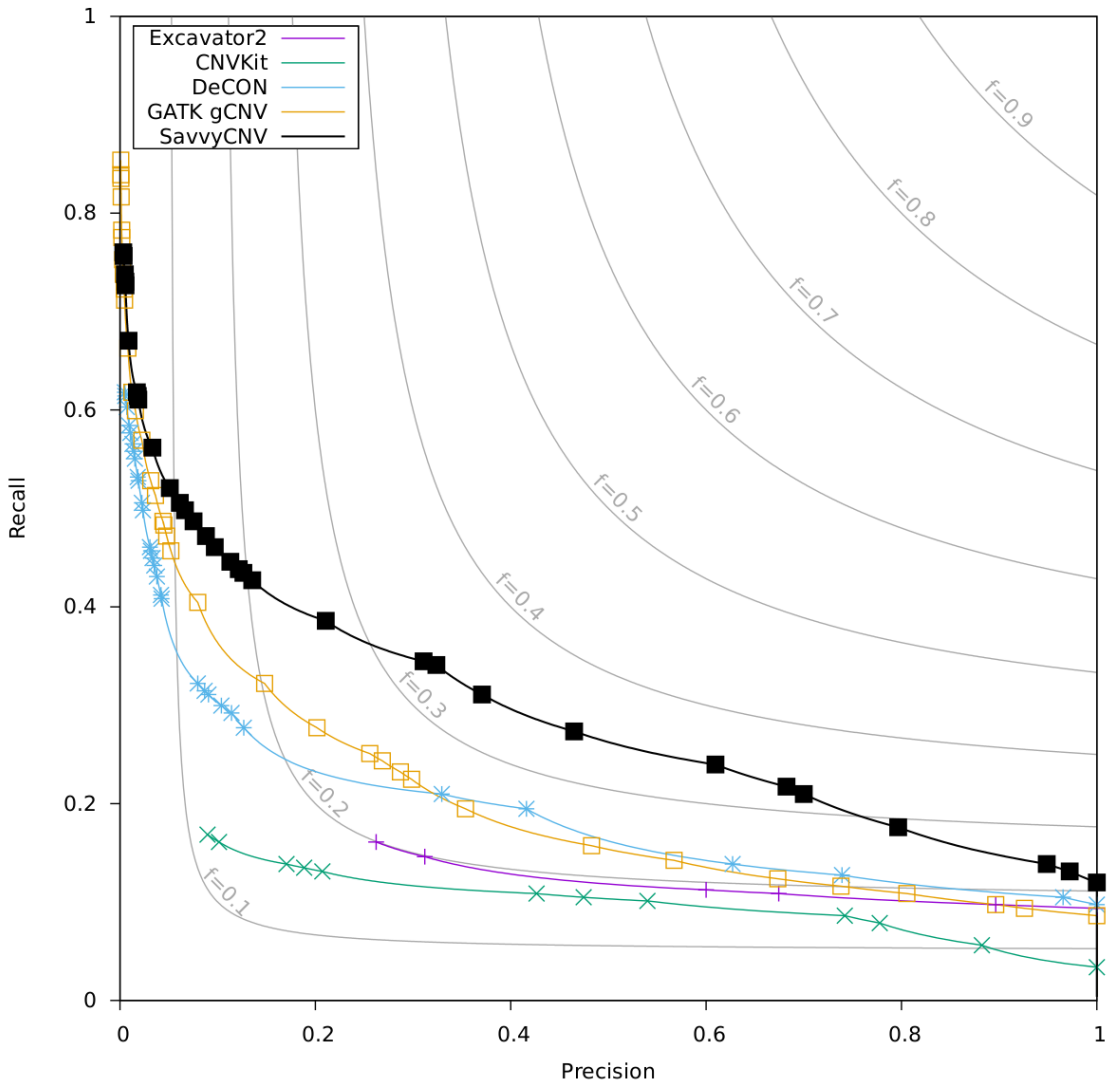

Figure 1. Benchmarking off-target CNV calling from targeted panel data.

The data points on the plot are generated by a parameter sweep for each tool and show the precision and recall that can be achieved with each tool. The f statistic is the harmonic mean of precision and recall (see methods for details).

88 **Table 1. Benchmarking off-target CNV calling from targeted panel data.**

89 **The table shows the performance of the different CNV calling software based on the size of the**
90 **CNV. The tools were run with multiple different parameters. For this comparison, we have**
91 **selected the configuration for each tool that provides the highest recall with a precision of at least**
92 **50%.**

Size	CNVs	Software	True positives	False positives	Recall	Precision
All	267	SavvyCNV	68	67	25.5%	50.4%
		GATK gCNV	40	31	14.9%	56.3%
		DeCON	41	40	15.4%	50.6%
		Excavator2	30	20	11.2%	60.0%
		CnvKit	27	23	10.1%	54.0%
>=1Mb	42	SavvyCNV	41	11	97.6%	78.8%
		GATK gCNV	39	32	92.9%	54.9%
		DeCON	36	31	85.7%	53.7%
		Excavator2	29	19	69.0%	60.4%
		CnvKit	26	24	61.9%	52.0%
>=5Mb	26	SavvyCNV	26	0	100%	100%
		GATK gCNV	26	3	100%	89.7%
		DeCON	26	19	100%	74.3%
		Excavator2	26	3	100%	89.7%
		CnvKit	25	23	96.2%	52.1%

93

94

95 All four tools called all of the CNVs larger than 5Mb, however only SavvyCNV did so without any false
96 positive calls. All CNVs larger than 1Mb were called by SavvyCNV, GATK gCNV, and DeCON, although
97 SavvyCNV called the most (97.6%) at a precision of above 50%. For all CNVs, SavvyCNV had the
98 highest recall (25.5%) with precision above 50%. For all three CNV size categories, SavvyCNV had the
99 greatest detection power. It can call CNVs that are larger than 1Mb from off-target reads from a
100 targeted panel with good recall (97.6%) and precision (78.8%).

101 **SavvyCNV can call on-target CNVs from targeted panels**

102 To evaluate the performance of SavvyCNV at calling CNVs from on-target data we used the ICR96
103 validation series(19). ICR96 is a set of 96 samples sequenced using a small targeted sequencing panel
104 (TruSight Cancer Panel v2, 100 genes), with exon CNVs detected independently using MLPA (25
105 single-exon CNVs, 43 multi-exon CNVs, and 1752 normal copy number genes). SavvyCNV had the
106 highest recall for precision over 50% though GATK gCNV and DeCON also performed well - all 3 tools
107 had a recall >95% (Table 2). Precision can only be compared between tools if recall is identical. While
108 GATK gCNV achieves 85.7% precision at its highest recall of 97.1%, SavvyCNV has a precision of
109 93.0% at the same recall (this is shown in Figure 2). DeCON was the next-best performing tool while
110 CnvKit did not call the majority of CNVs. Excavator2 did not run on this data set. Figure 2 shows the
111 recall and precision of the four tools. SavvyCNV was the only tool capable of detecting all the CNVs
112 although only with a precision of 29.1%.

113

114

115

116

117

118 **Table 2 Benchmarking on-target CNV calling from the ICR96 targeted panel data.**

119 **The table shows the performance of the different CNV calling software based on the size of the**
120 **CNV. The tools were run with multiple different parameters. For this comparison, we have**
121 **selected the configuration for each tool that provides the highest recall with a precision of at least**
122 **50%.**

Size	CNVs	Software	True positives	False positives	Recall	Precision
All	**68**	SavvyCNV	67	42	98.5%	61.5%
		GATK gCNV	66	11	97.1%	85.7%
		DeCON	65	65	95.6%	50.0%
		CnvKit	12	12	17.6%	50.0%
Multi-exon	**43**	SavvyCNV	43	11	100%	79.6%
		GATK gCNV	43	11	100%	79.6%
		DeCON	42	14	97.7%	75.0%
		CnvKit	10	8	23.2%	55.5%

123

124

125

126

127

128

129 Two of the CNVs within the ICR96 dataset cover less than a complete exon and have one breakpoint
130 within the targeted region. These two CNVs are the hardest to detect by read-depth methods, as the

read depth is only altered over a fraction of the exon area. Both CNVs are detected only by SavvyCNV, even when the highest sensitivity settings are used with the other CNV callers.

Multi-exon CNVs are easier to detect than single-exon CNVs. SavvyCNV, GATK gCNV, and DeCON can detect all 43 multi-exon CNVs, although only SavvyCNV and GATK gCNV did this with a precision greater than 50%.

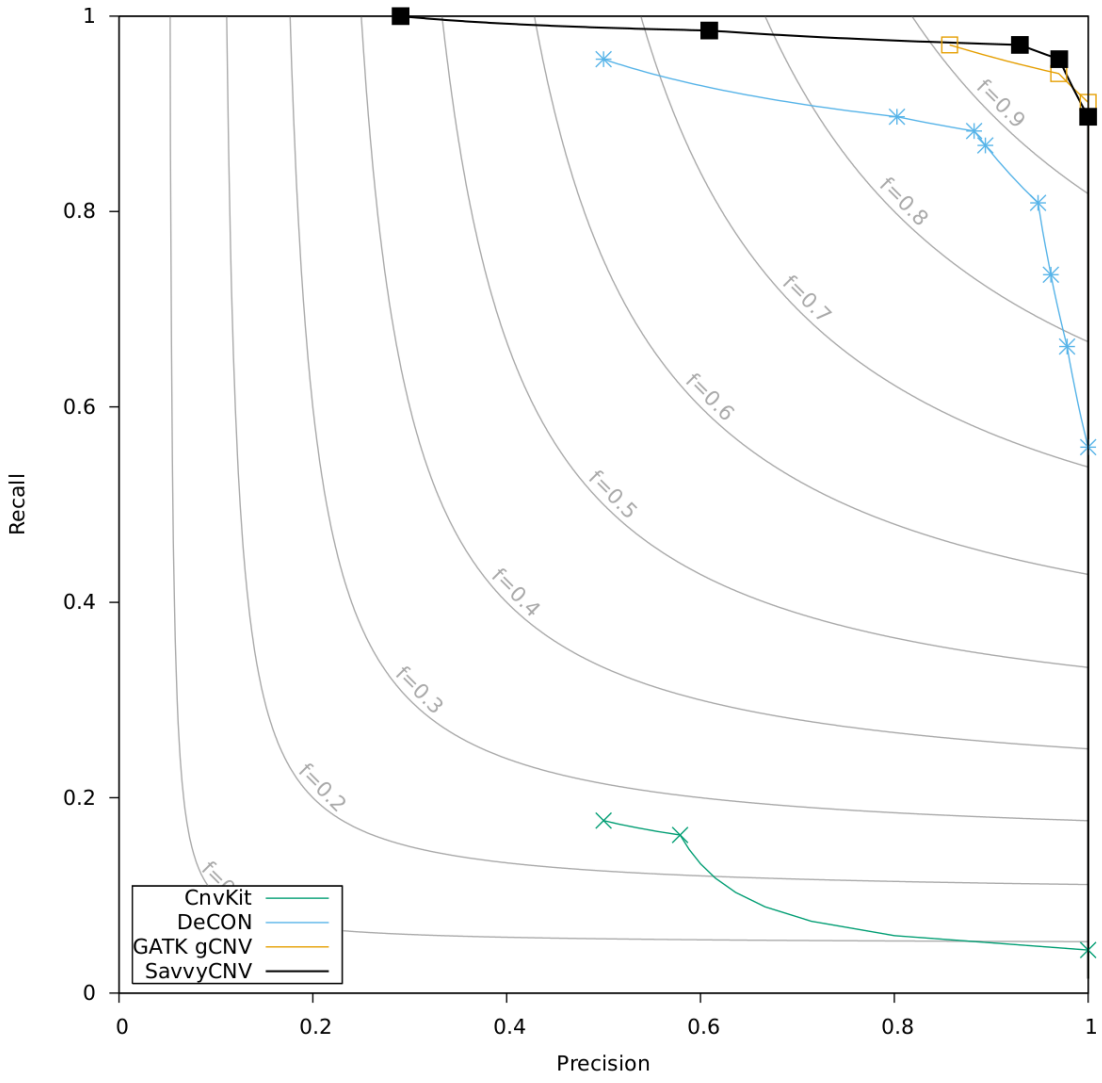

Figure 2. Benchmarking on-target CNV calling from targeted panel data.

The data points on the plot are generated by a parameter sweep for each tool and show the precision and recall that can be achieved with each tool. The f statistic is the harmonic mean of precision and recall (see methods for details).

SavvyCNV can call off-target CNVs from exome data

To assess SavvyCNV's ability to call CNVs from off-target reads generated by exome sequencing we benchmarked it against a truth set (see Methods) and compared its performance to GATK gCNV, DeCON, EXCAVATOR2, and CNVkit. The best recall where precision is at least 50% is shown in Table 3 for two different size categories, and recall/precision is shown in Figure 3 for all CNVs.

SavvyCNV was the best performing tool on this data set, able to call 84% of the CNVs with 50% precision, while the next best tool (DeCON) called 38.7% of CNVs with 50% precision. The chief difference between the performances of the tools is SavvyCNV's ability to call CNVs smaller than 200kb. SavvyCNV is able to call an additional 29 CNVs that are smaller than 200kB at >50% precision while GATK gCNV, EXCAVATOR2, and CNVkit call no true CNVs smaller than 200kB and DeCON calls 7.

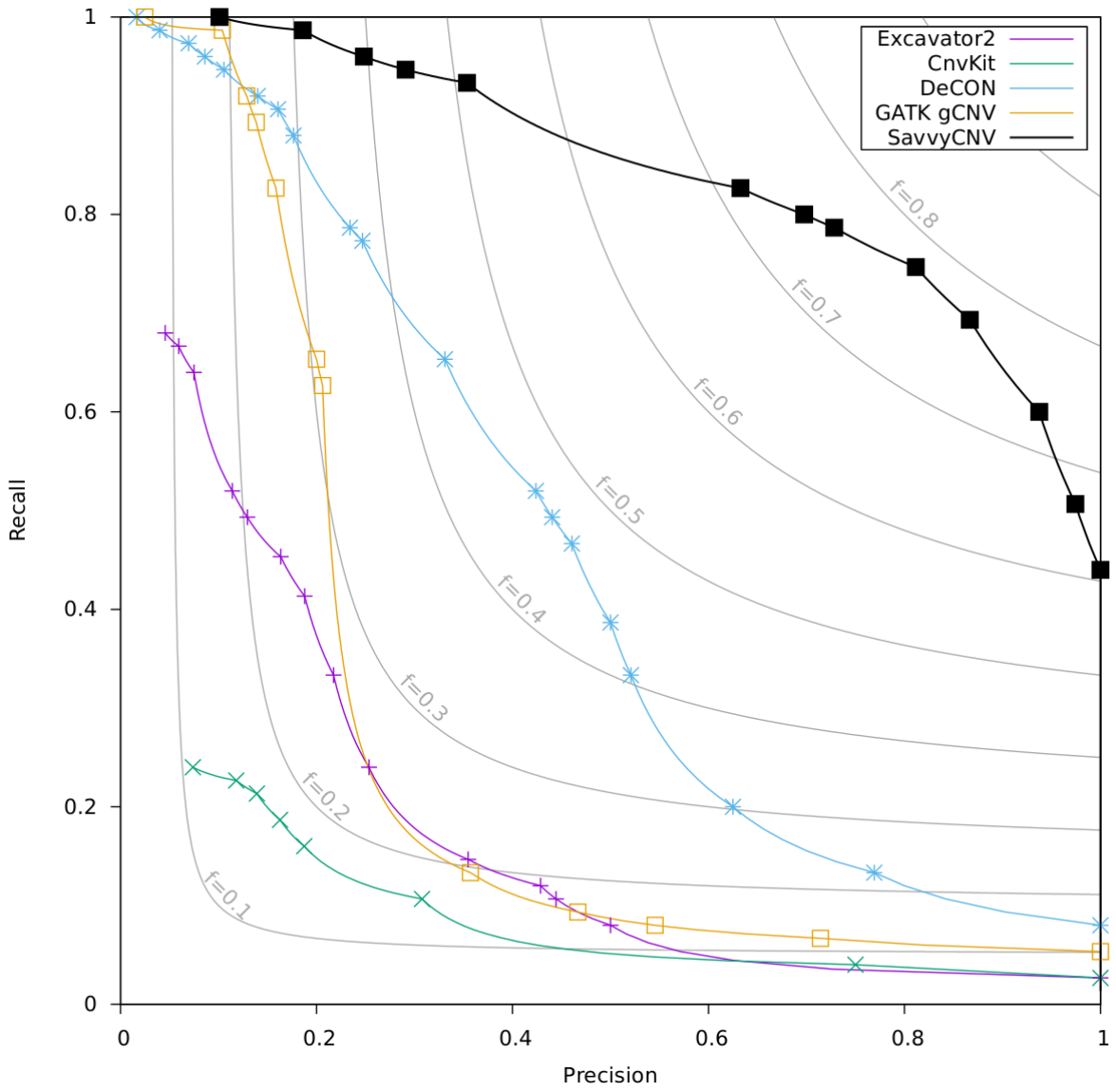

Figure 3. Benchmarking off-target CNV calling from exome data.

The data points on the plot are generated by a parameter sweep for each tool and show the precision and recall that can be achieved with each tool. The f statistic is the harmonic mean of precision and recall (see methods for details).

158 **Table 3. Benchmarking on-target CNV calling from the exome data.**

159 **The table shows the performance of the different CNV calling software based on the size of the**
160 **CNV. The tools were run with multiple different parameters. For this comparison, we have**
161 **selected the configuration for each tool that provides the highest recall with a precision of at least**
162 **50%.**

Size	CNVs	Software	True positives	False positives	Recall	Precision
All	75	SavvyCNV	63	56	84.0%	52.9%
		GATK gCNV	6	5	8.0%	54.5%
		DeCON	29	29	38.7%	50.0%
		Excavator2	6	6	8.0%	50.0%
		CnvKit	3	1	4.0%	75.0%
>=200kb	40	SavvyCNV	34	27	85.0%	55.7%
		GATK gCNV	6	5	15.0%	54.5%
		DeCON	22	20	55.0%	52.4%
		Excavator2	6	6	15.0%	50.0%
		CnvKit	3	1	7.5%	75.0%

163

164

165

166

167

168

169

170 **SavvyCNV can detect clinically relevant CNVs**

171 Having validated the ability of SavvyCNV to call CNVs from off-target reads we proceeded to screen
172 for CNVs in our full cohort of targeted panel samples from patients referred for genetic testing to
173 identify the cause of their diabetes or hyperinsulinism(16). We were able to detect 11 clinically
174 relevant CNVs both within and outside of the targeted regions (Table 4). Of these, 4 provided a new
175 genetic diagnosis for diabetes/hyperinsulinism (rows 1-4 in table 4), providing information which will
176 guide clinical management and allow accurate counselling on recurrence risk in family members and
177 future offspring. The remaining 7 CNVs (rows 5-11 in table 4) confirmed clinically-reported
178 diagnoses unrelated to the diabetes/hyperinsulinism. These findings demonstrate the ability of
179 SavvyCNV to detect clinically relevant CNVs and aneuploidies from off-target data from a small
180 targeted panel.

181 **Table 4 – Clinically-relevant CNVs detected.**

Row	CNV detected	CNV size	Clinical confirmation	Reason for referral	Clinical implications
1	Chr8:6,800,000-11,800,000 deletion	5Mb	Deletion includes *GATA4*, causative of the patient's neonatal diabetes and additional features.	Diabetes	Genetic diagnosis of monogenic diabetes (20).
2	Chr8:8,000,000-10,400,000 duplication 8:10,600,000-	2.4Mb and 1.4Mb	Deletion includes *GATA4*, causative of the patient's neonatal diabetes and additional features.	Diabetes	Genetic diagnosis of monogenic diabetes (20).

	12,000,000 deletion				
3	Chr18:19,400,000-21,800,000 deletion	2.4Mb	Deletion includes *GATA6*, causative of the patient's neonatal diabetes and additional features.	Diabetes	Genetic diagnosis of monogenic diabetes (21).
4	ChrX:0-57,000,000 deletion X:76,200,000-155,400,000 deletion	57Mb and 79.2Mb	Ring X chromosome confirmed by cytogenetics.	Hyperinsulinism	Genetic diagnosis of Turner syndrome, a known cause of hyperinsulinism (22).
5	Chr21:9,400,000-48,200,000 duplication	Chromosome	Patient known to have Down Syndrome at referral.	Hyperinsulinism	Confirms the diagnosis of Down syndrome.
6	Chr21:10,000,000-48,200,000 duplication	Chromosome	Patient known to have Down Syndrome at referral.	Diabetes	Confirms the diagnosis of Down syndrome (23).
7	Chr21:11,000,000-48,200,000 duplication	Chromosome	Patient known to have Down Syndrome at referral.	Hyperinsulinism	Confirms the diagnosis of Down syndrome.
8	Chr21:14,400,000-	Chromosome	Patient known to have Down Syndrome at	Hyperinsulinism	Confirms the diagnosis of Down

	48,200,000 duplication		referral.		syndrome.
9	Chr22:18,800,000-21,600,000 deletion	1.8Mb	Confirmed by array CGH.	Diabetes	Provided the diagnosis of DiGeorge syndrome.
10	Chr22:18,800,000-21,600,000 deletion	1.8Mb	Patient known to have DiGeorge syndrome at referral.	Hyperinsulinism	Confirms the diagnosis of DiGeorge syndrome.
11	ChrX:0-155,400,000 duplication	Chromosome	Patient known to have XXX syndrome at referral.	Diabetes	Confirms the diagnosis of XXX syndrome.

Discussion

SavvyCNV can detect CNVs genome-wide from off-target reads

We benchmarked SavvyCNV on its ability to call off-target and on-target CNVs from targeted panel and exome sequencing data. This new tool outperformed four existing tools, two of which (CNVkit(15) and Excavator2 (14)) were specifically designed to call off-target CNVs. GATK gCNV performed similarly to SavvyCNV in the on-target (ICR96) analysis. However, SavvyCNV considerably outperforms all other tools in the off-target analyses.

SavvyCNV finds the greatest number of true positive CNVs in all data sets while other tools did not call certain CNVs. For example, the two partial exon CNVs in the on-target (ICR96) data set are detected only by SavvyCNV. This is likely because of the improved error correction and error

193 modelling that is incorporated into SavvyCNV over existing tools. SavvyCNV uses singular vector
194 decomposition to reduce noise. CNVkit and EXCAVATOR2 only correct for GC content, while GATK
195 gCNV uses Bayesian principle component analysis (24), and DeCON uses sample matching (it
196 searches for samples in the control set that have a similar noise profile). Supplementary figure 1
197 demonstrates how error correction improves the recall and precision. Supplementary figure 2 shows
198 that the bespoke error model used by SavvyCNV (see supplementary methods) performed better
199 than the Poisson error model used by the other tools, as this uses the error information available
200 from having multiple control samples.

201 SavvyCNV had a higher precision than other tools when calling off-target CNVs, an important
202 consideration in diagnostic and research laboratories as if false positives are reduced, fewer CNVs
203 will require orthogonal testing to identify the true positive results. Many of the false positives
204 produced by DeCON, CNVKit, and EXCAVATOR2 have a read depth ratio indicating that they are
205 either mosaic CNVs or random noise. The prior probability is overwhelmingly that these are random
206 noise. This is why the default for SavvyCNV and GATK gCNV is to call only non-mosaic CNVs as this
207 hugely reduces the number of false positives called. Mosaic CNV calling can be enabled in SavvyCNV
208 for projects where it is applicable. Supplementary figure 3 demonstrates the improvement in
209 precision of the default mode compared to the mode that includes calling of mosaic CNVs.

210 **Estimates of precision and recall rely on the quality of the truth set**

211 On-target CNV calling from a targeted panel was tested on the ICR96 data set in which the truth set
212 was verified by MLPA. The truth sets for the off-target CNV calling from targeted panel and exome
213 sequence data were generated from CNV calls from genome sequencing data. Genome sequencing
214 has a much higher coverage than that generated from only off-target reads which allows CNVs to be
215 called more accurately enabling them to be used as a truth set. GenomeStrip(25) was used to call
216 the truth set as it was designed to call CNVs from genome data and was not one of the tools under
217 examination in this study. However, it is possible that there could be some false positive and

218 negative calls in the truth set. This would lower the precision and recall of the tools under
219 examination but should not bias the results in favour of a particular tool.

220 **Sensitivity depends on the size of the CNV**

221 Smaller CNVs are harder for all software to detect. For all tools tested the larger the CNV the better
222 the precision and recall, however SavvyCNV performs better than the other tools tested. SavvyCNV
223 detects CNVs above 1Mb with 100% recall in off-target data from both targeted panel and exome
224 data.

225 **CNV calling can be optimised for precision or recall by adjusting configuration**

226 When calling CNVs, precision and recall are a trade-off; high recall will maximise the number of true
227 CNVs that are called, with the consequence that it also reduces precision resulting in a large number
228 of false positive CNV calls. Different precision levels are appropriate in different situations,
229 influenced both by the experimental methodology and the aims of the project. When calling CNVs
230 on-target on a small gene panel there will be fewer false positive calls generated due to the smaller
231 target area thus it may be preferable to adjust settings to enable a higher recall at the cost of a
232 lower precision. This could also be true in a clinical context where the most important aim is to not
233 miss a true causative variant. In contrast, when calling CNVs genome-wide in a gene-agnostic
234 approach such as genome sequencing, a higher precision is likely to be desirable to avoid generating
235 an unmanageably long list of CNVs. The user can choose their preferred settings for SavvyCNV for
236 different project requirements.

237 **Off-target CNV calling is 'free' data and increases diagnostic yield**

238 SavvyCNV utilises data already generated by targeted panel and exome tests. These tests are carried
239 out in order to detect single nucleotide variants and small insertions or deletions (<50 base pairs). In
240 some laboratories CNVs are also detected within the targeted regions using CNV calling software
241 while other laboratories use array-CGH or MLPA to detect CNVs. Using SavvyCNV allows CNVs to be

242 detected not just within the targeted regions but allows genome-wide CNV calling. This will provide
243 a genetic diagnosis for more patients, increasing the diagnostic yield of these tests. We have
244 demonstrated the ability to find relevant genetic diagnoses using off-target CNV calling from our
245 small targeted panel. Existing data can be reanalysed with our method to reveal additional CNVs. As
246 an illustration of this, two of the CNVs in the ICR96 data set were found to actually be large CNVs
247 (15Mb and 56Mb), which may have clinical implications beyond the targeted gene.

248 Conclusions

249 SavvyCNV calls CNVs from off-target reads from exomes and small targeted gene panels with high
250 precision and recall, and performs better than existing tools including those designed for off-target
251 CNV calling. Calling CNVs from off-target reads is exploiting 'free' data to increase the diagnostic
252 yield of targeted panel and exome sequencing tests and reveal important biological findings.

253 Methods

254 Targeted panel data

255 2591 patients were referred to the molecular genetics department at the Royal Devon and Exeter
256 Hospital for genetic testing for maturity onset diabetes of the young (MODY), neonatal diabetes
257 (NDM) or hyperinsulinemic hypoglycemia (HH). Informed consent was obtained from the probands
258 or their parents/guardians and the study was approved by the North Wales ethics committee.
259 Samples were sequenced on a targeted gene panel test for monogenic diabetes and HH using a
260 custom Agilent SureSelect panel of 75 genes, targeting 200kb and obtaining 3.4 million reads on
261 average per sample (standard deviation 1.6 million)(16), using an Illumina HiSeq 2500 or an Illumina
262 NextSeq 500. Based on the GATK best practice guidelines(17) reads were aligned to the
263 hg19/GRCh37 human reference genome with BWA mem(26), duplicates were removed using Picard
264 (27) and GATK IndelRealigner was used for local re-alignment (28). CNV analysis was carried out on
265 these BAM files.

266 **Exome sequencing data**

267 Following testing using the targeted panel above, samples from 86 patients underwent exome
268 sequencing with Agilent SureSelect Whole Exome versions 1, 3, 4, and 5, obtaining 76 million reads
269 on average (standard deviation 20 million). Sequencing, alignment and variant calling was as above
270 for the targeted panel.

271 **Truth set for targeted and exome data**

272 170 of the targeted panel samples and 42 of the exome samples were subsequently genome
273 sequenced on an Illumina HiSeq 2500 or an Illumina HiSeq X10. These were used to create a truth
274 set of CNVs for testing the off-target CNV calling from targeted panel or exome data. The CNVs in the
275 truth set were called by GenomeStrip(25) from the genome sequencing data. In order to remove
276 false positive calls CNVs were filtered based on their allele balance ratios – whether the allele
277 balance of the variants within the called CNV was consistent with it being a true call. We used the X
278 chromosome in males to calibrate the expected allele ratio for a deletion and used the allele ratio of
279 normal, two copy regions to evaluate if the allele ratio for duplications fell above that. In addition 37
280 CNVs were added to the targeted panel truth set as they were validated by other methods such as
281 Multiplex Ligation-dependent Probe Amplification (MLPA) or were aneuploidies reported by the
282 clinician at time of referral for genetic testing.

283 The ICR96 data set (19) was used to benchmark on-target CNV calling. This data set consists of 96
284 samples sequenced on a targeted panel where the truth set of CNVs is based on 68 positive and
285 1752 negative MLPA tests.

286 **Calling clinically-relevant CNVs**

287 The remaining 2479 targeted panel samples from unsolved patients with MODY, NDM and HH were
288 analysed with SavvyCNV to look for off-target CNVs which might explain their phenotype. For clinical
289 evidence of the CNVs, see Table 4.

290 **CNV tool comparisons**

291 To ensure a fair comparison between the different tools, for each data set all tools were run with a
292 variety of configurations. The size of genomic regions that were analysed was varied for all five tools
293 (targeted panel: 150,000bp to 300,000bp; exomes: 6,000bp to 50,000bp; ICR96: 200bp to 600bp).
294 The hidden Markov model transition probability was varied for DeCON and SavvyCNV (10^{-10} to 0.1).
295 All five tools provide quality metrics for the CNV calls. These metrics were used to filter the CNV calls
296 to reject false positive calls and retain true positive calls. All possible quality cut-off values were
297 tried. The best precision achieved for each possible recall was then selected for each tool from all
298 the generated results, and plotted in precision-recall graphs. EXCAVATOR2(14) did not run on the
299 ICR96 dataset - we contacted the authors of the tool but did not receive a response.

300 **Statistics**

301 We defined recall as the percentage of true positive CNVs that were found by the tool. We defined
302 precision as the percentage of the total CNVs called by the tool that were true. Several figures use
303 the f statistic to compare tools; this is the harmonic mean of precision and recall.

304 List of abbreviations

305 CNV: Copy Number Variant

306 NGS: Next Generation Sequencing

307 MLPA: Multiplex Ligation-dependent Probe Amplification

308 Declarations

309 **Ethics**

310 The study was conducted in accordance with the Declaration of Helsinki.

311 Consent for publication – not applicable.

312 **Availability of data and material**

313 SavvyCNV is available for download from https://github.com/rdemolgen/SavvySuite

314 The targeted panel and exome sequencing data analysed during the current study is not publicly
315 available due to patient confidentially.

316 The ICR96 data set is publically available(19)

317 **Competing interests**

318 The authors declare that they have no competing interests

319 **Funding**

320 SE is the recipient of a Wellcome Senior Investigator award (grant number WT098395/Z/12/Z). SEF
321 has a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (grant
322 number: 105636/Z/14/Z). KAP holds a post-doctoral fellowship from the Wellcome trust
323 (110082/Z/15/Z). This work was undertaken as part of the Transforming Genetic Medicine Initiative
324 funded by the Wellcome Trust.

325 The funding bodies had no role in the design of the study or the collection, analysis or interpretation
326 of data.

327 **Authors' contributions**

328 SE, SEF, TWL, MNWe and MNWa designed the study. SEF, EDF, MJ, KP and SE collected and analysed
329 patient data. MNWa wrote the software. MNWa and TWL wrote the manuscript. All authors
330 approved the final version of the manuscript.

331 **Acknowledgements**

The authors would like to thank Professor Andrew Hattersley, Dr Jayne Houghton and Andrew Parish (Department of Molecular Genetics, Royal Devon & Exeter NHS Foundation Trust, Exeter, UK) for their technical assistance and advice.

References

1. The European Polycystic Kidney Disease C. The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. Cell. 1994;77(6):881-94.
2. Gardner R, Mungall A, Dunham I, Barber J, Shield J, Temple I, et al. Localisation of a gene for transient neonatal diabetes mellitus to an 18.72 cR(3000) (~5.4 Mb) interval on chromosome 6q. J Med Genet. 1999;36(3):192-6.
3. Macé A, Tuke MA, Deelen P, Kristiansson K, Mattsson H, Nõukas M, et al. CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. Nature Communications. 2017;8:744.
4. Gaedigk A, Blum M, Gaedigk R, Eichelbaum M, Meyer UA. Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. Am J Hum Genet. 1991;48(5):943-50.
5. Smith MJ, Urquhart JE, Harkness EF, Miles EK, Bowers NL, Byers HJ, et al. The Contribution of Whole Gene Deletions and Large Rearrangements to the Mutation Spectrum in Inherited Tumor Predisposing Syndromes. Hum Mutat. 2016;37(3):250-6.
6. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong Association of De Novo Copy Number Mutations with Autism. Science (New York, NY). 2007;316(5823):445-9.
7. Solomon BD, Rosenbaum KN, Meck JM, Muenke M. Holoprosencephaly due to Numeric Chromosome Abnormalities. American journal of medical genetics Part C, Seminars in medical genetics. 2010;154C(1):146-8.
8. Fakhro KA, Choi M, Ware SM, Belmont JW, Towbin JA, Lifton RP, et al. Rare copy number variations in congenital heart disease patients identify unique genes in left-right patterning. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(7):2915-20.
9. Docherty LE, Kabwama S, Lehmann A, Hawke E, Harrison L, Flanagan SE, et al. Clinical presentation of 6q24 transient neonatal diabetes mellitus (6q24 TNDM) and genotype–phenotype correlation in an international cohort of patients. Diabetologia. 2013;56(4):758-62.
10. Antonarakis SE, Lyle R, Dermitzakis ET, Reymond A, Deutsch S. Chromosome 21 and Down syndrome: from genomics to pathophysiology. Nature Reviews Genetics. 2004;5:725.
11. Aten E, White SJ, Kalf ME, Vossen RH, Thygesen HH, Ruivenkamp CA, et al. Methods to detect CNVs in the human genome. Cytogenet Genome Res. 2008;123(1-4):313-21.
12. Samarakoon PS, Sorte HS, Kristiansen BE, Skodje T, Sheng Y, Tjønnfjord GE, et al. Identification of copy number variants from exome sequence data. BMC Genomics. 2014;15(1):661.
13. Samuels DC, Han L, Li J, Quanghu S, Clark TA, Shyr Y, et al. Finding the lost treasures in exome sequencing data. Trends in genetics : TIG. 2013;29(10):593-9.
14. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Res. 2016;44(20):e154-e.
15. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Computational Biology. 2016;12(4):e1004873.

16. Ellard S, Lango Allen H, De Franco E, Flanagan SE, Hysenaj G, Colclough K, et al. Improved genetic testing for monogenic diabetes using targeted next-generation sequencing. Diabetologia. 2013;56(9):1958-63.
17. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297-303.
18. Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Research. 2016;1:20.
19. Mahamdallie S, Ruark E, Yost S, Ramsay E, Uddin I, Wylie H, et al. The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. Wellcome Open Research. 2017;2:35.
20. Shaw-Smith C, De Franco E, Lango Allen H, Batlle M, Flanagan SE, Borowiec M, et al. GATA4 Mutations Are a Cause of Neonatal and Childhood-Onset Diabetes. Diabetes. 2014;63(8):2888-94.
21. De Franco E, Shaw-Smith C, Flanagan SE, Shepherd MH, Hattersley AT, Ellard S. GATA6 Mutations Cause a Broad Phenotypic Spectrum of Diabetes From Pancreatic Agenesis to Adult-Onset Diabetes Without Exocrine Insufficiency. Diabetes. 2013;62(3):993-7.
22. Gibson CE, Boodhansingh KE, Li C, Conlin L, Chen P, Becker SA, et al. Congenital Hyperinsulinism in Infants with Turner Syndrome: Possible Association with Monosomy X and KDM6A Haploinsufficiency. Horm Res Paediatr. 2018:413-22.
23. Johnson MB, De Franco E, Atma W Greeley S, Letourneau LR, Gillespie K, , et al. Trisomy 21 is a Cause of Permanent Neonatal Diabetes that is Autoimmune but not HLA Associated. Diabetes. 2019.
24. BroadInstitute. A scalable Bayesian framework for inferring copy number variation 2019 [Available from: https://www.broadinstitute.org/videos/scalable-bayesian-model-copy-number-variation-bayesian-pca.
25. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. Nature Genetics. 2015;47:296.
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
27. BroadInstitute. Picard [Available from: http://broadinstitute.github.io/picard/.
28. Van der Auwera G. IndelRealigner [Available from: https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php.