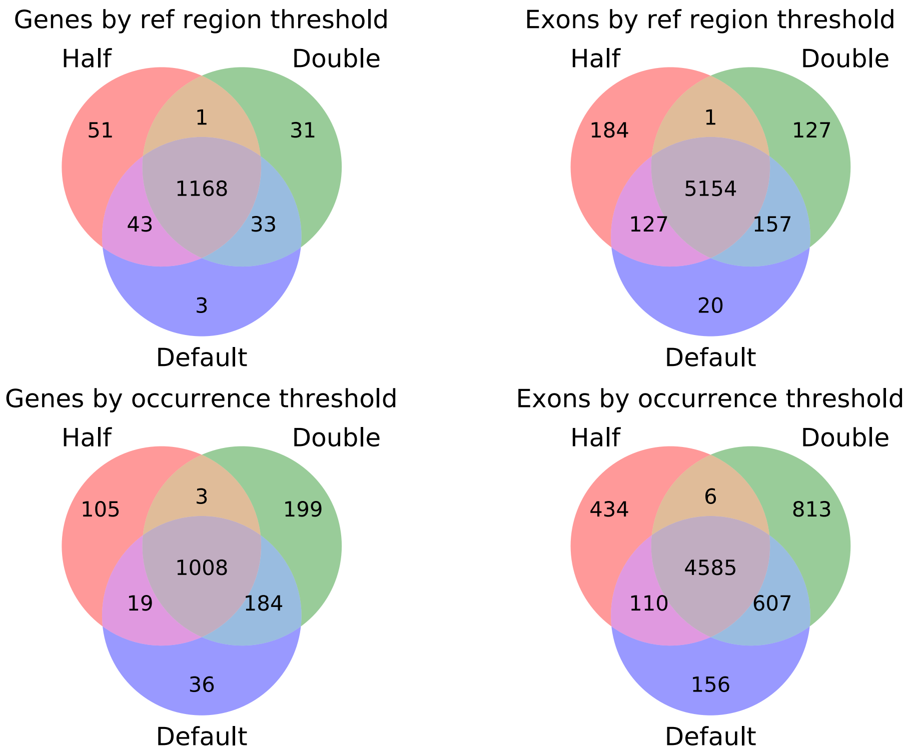
Supplementary Data for manuscript:

WISExome: A within-sample comparison approach to detect copy number variations in whole exome sequencing data.

Roy Straver[1], Marjan M. Weiss[1], Quinten Waisfisz[1], Erik A. Sistermans[1,*] and Marcel J.T. Reinders[1,2,*]

ID		Array				WISExome			XHMM			CoNIFER			CODEX			CLAMMS		
CNV	Sample	Chr	Start	Stop	Amp/Del	Start	Stop	Calls	Start	Stop	Calls	Start	Stop	Calls	Start	Stop	Calls	Start	Stop	Calls
1	A	15	23605900	28771100	Del	23608251	28595888	1	23609709	28566641	49	23608042	25420097	1	23577463	28632880	4	23631315	28525469	1
2	B	9	214187	4566187	Amp	64174	4576676	1	271660	4564531	43	35404	4583226	1	154537	4585524	2	214478	4585524	1
3	B	7	100000	2818074	Del	193521	2831309	1	288235	2802433	48	-	-	0	20781	2802433	4	195510	2802536	1
4	C	15	31073761	32915593	Del	30902878	32446230	1	31114786	32404147	12	-	-	0	30875082	32691203	2	31085444	32404585	1
5	D	8	8774423	10475458	Amp	8750463	10411575	1	8877786	10396230	12	-	-	0	8750463	10411575	1	8865536	10411575	1
6	E	X	6448777	8134649	Amp	6451945	8095152	1	6451946	8095152	5	-	-	0	6968286	8095152	1	6968286	7894192	1
7	F	20	540200	2208991	Del	590255	2129263	1	642660	2129263	17	-	-	0	585164	2129263	1	585164	2129263	1
8	G	22	20723711	21951312	Amp	20715521	21663254	2	20724361	21424388	22	-	-	0	20705764	21537878	2	20754816	21424388	1
9	H	6	145551394	146283017	Amp	145669988	146276201	1	145956422	146276015	15	145172405	146350806	1	145669728	146276201	1	145669728	146276201	1
10	I	7	6010735	6361936	Amp	6022386	6414453	1	6026362	6217654	8	-	-	0	6017156	6230177	1	6026438	6230177	1
11	J	1	235422065	235716261	Amp	235423872	235658195	1	235490245	235643528	11	-	-	0	235423872	235658195	1	235423872	235658195	1
12	K	21	27252861	27543138	Amp	27253167	27840939	1	27253168	27512571	6	-	-	0	27253167	27512571	1	27253167	27512571	1
13	L	16	28732295	28952277	Del	28769968	29062277	1	-	-	0	-	-	0	28734541	29050078	1	28836587	29001379	1
14	J	1	237807486	238000883	Amp	237811669	237997180	1	237811670	237996523	11	237806576	238037501	1	237811669	237997180	1	237811669	237997180	1
15	M	3	2818074	2980918	Del	2777518	2967517	1	-	-	0	-	-	0	2777518	2967517	1	2777679	2967517	1
16	N	9	107459905	107545550	Del	107456603	107558300	1	107528604	107543399	2	-	-	0	107456603	107558300	1	107456603	107556798	1
17	I	16	89813821	89864352	Del	89813160	89862474	1	89842085	89842279	1	-	-	0	89813160	89862474	1	89813160	89862474	1
18	O	5	140222138	140238656	Del	140222881	140238082	1	140222088	140238082	5	140221670	140242395	1	140230761	140238082	1	-	-	0
19	P	X	153058879	153068482	Del	153059628	153074113	1	-	-	0	-	-	0	153059944	153074113	1	153059944	153074113	1
20	Q	5	7867073	7871862	Del	7866980	7871059	1	7869175	7871059	1	-	-	0	7867119	7871059	1	7867119	7871059	1

Table S1: Overview of all known pathogenic CNVs in 17 test samples. The first two columns indicate CNV index and sample identifier. The array based information is shown in the next four columns, showing the chromosome (Chr), start and stop positions (based on Hg19), and whether the call was an amplification or a deletion (Amp/Del). Next, for every tool the start and end positions for calls overlapping the array calls are shown, as well as the number of calls that overlapped. Whenever multiple calls were made, the reported start position is the first base pair position of the first call with overlap, the stop position is the last base pair of the last call overlapping the array CNV, and the number of overlapping calls Is shown in the corresponding Calls column.


Genes by ref region threshold

Exons by ref region threshold

Genes by occurrence threshold

Exons by occurrence threshold

Figure S1: Venn diagram of overlapping genes (left) and exons (right) when running WISExome with three settings of pre-filtering of target regions. WISExome's default setting is: pruning target regions with less than 10 reference target regions, as well as target regions that are called in more 4 training samples. Top row shows two different settings for the number of reference targets: (1) HALF: prune target regions with less than 5 reference target regions, and (2) DOUBLE: prune regions with less than 20 reference target regions. The bottom row shows two different settings for the number of calls in the training set: (1) HALF: prune target regions that occur more than 2 times in the training set, (2) DOUBLE: prune regions that occur more than 8 times in the training set.

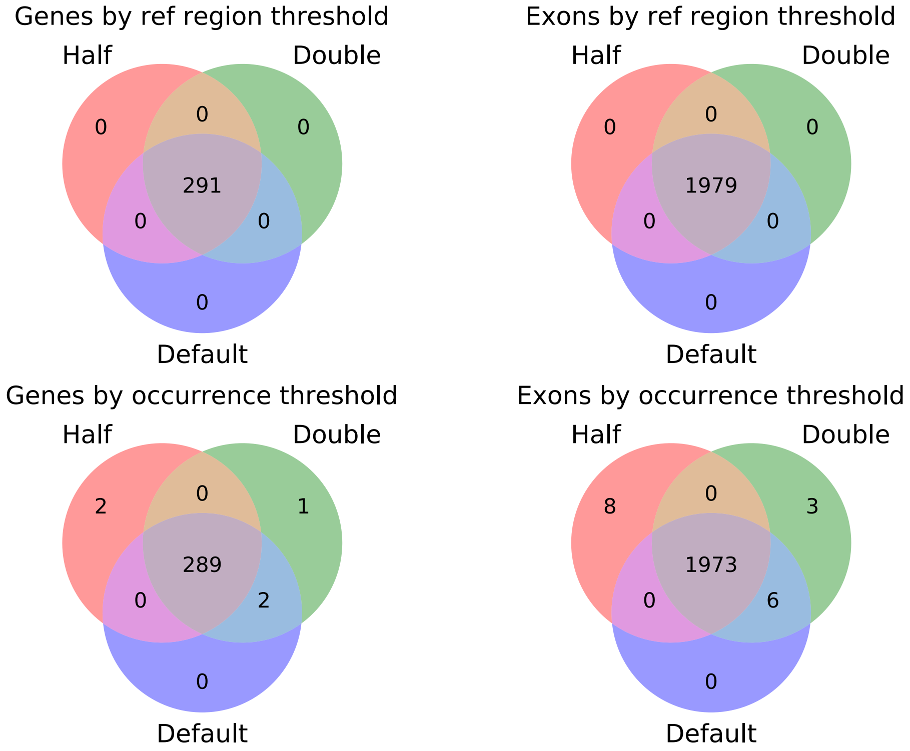


Figure S2: Similar to Figure S1, showing the effect of two of the pre-filtering thresholds. Here, the calls of WISExome for all the different setting are first filtered to require overlap with the regions detected by the array analysis. This way we can inspect how the different filter settings influence results that overlap with the array analysis.

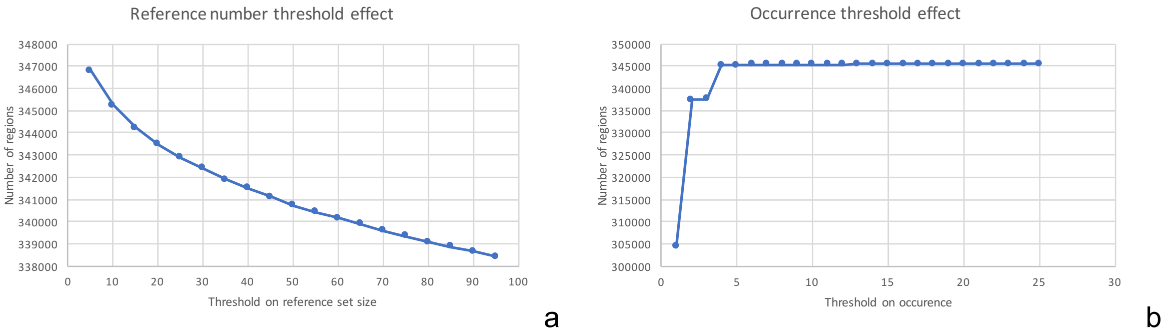
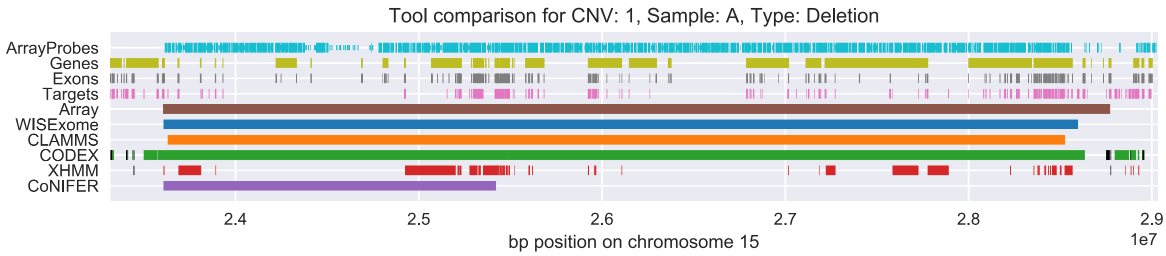
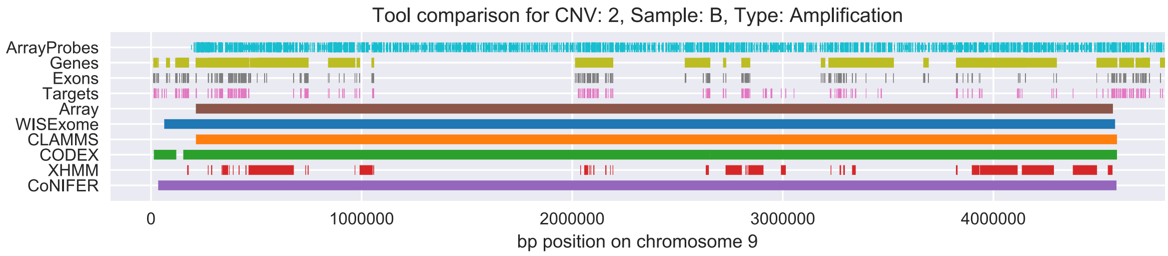
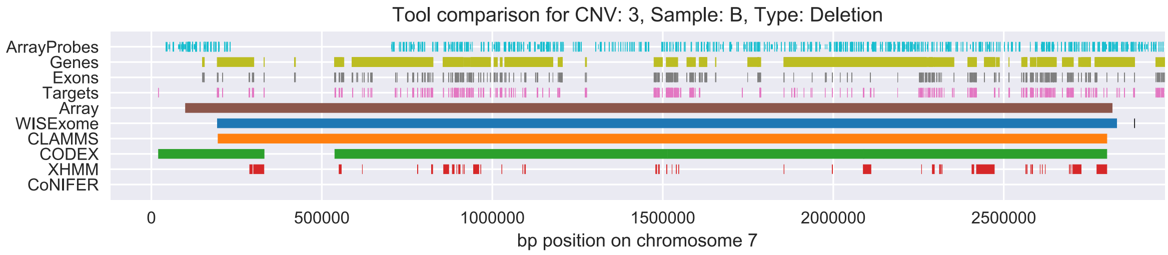
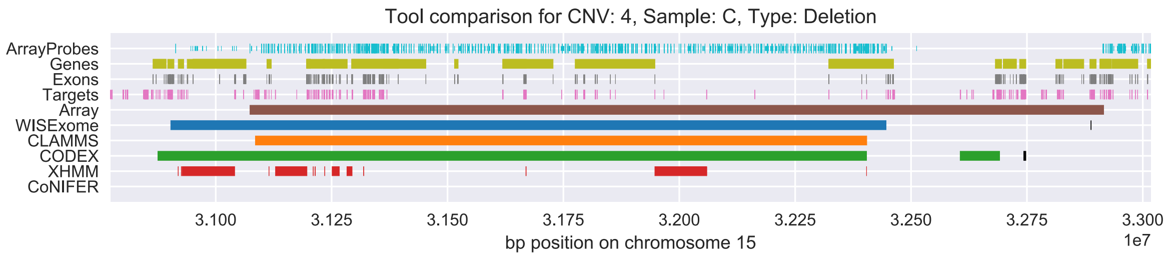


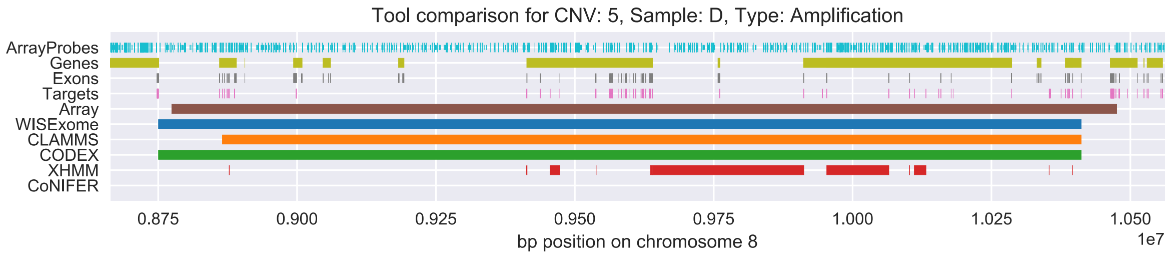
Figure S3: Effect of thresholds on the minimum number of reference target regions (a) and the maximal number of calls within the training set (b) on the number of reliable target regions (vertical axis).

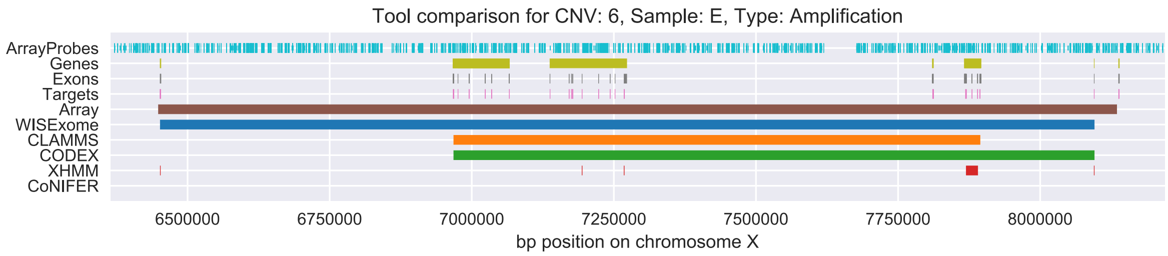

Tool comparison for CNV: 1, Sample: A, Type: Deletion
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
2.4
2.5
2.6
2.7
2.8
2.9
1e7
bp position on chromosome 15

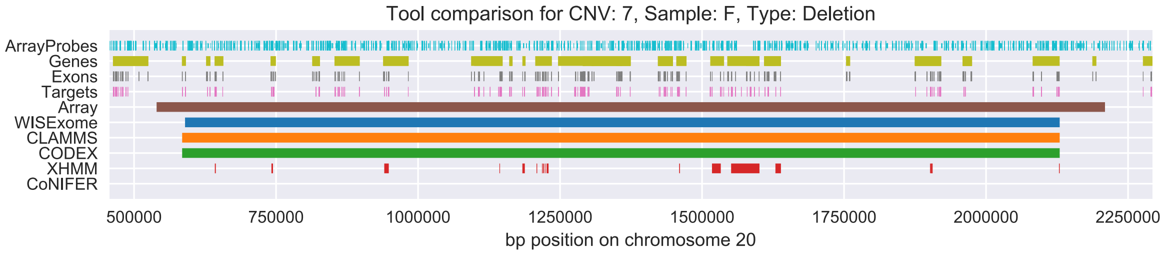

Tool comparison for CNV: 2, Sample: B, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
0
1000000
2000000
3000000
4000000
bp position on chromosome 9

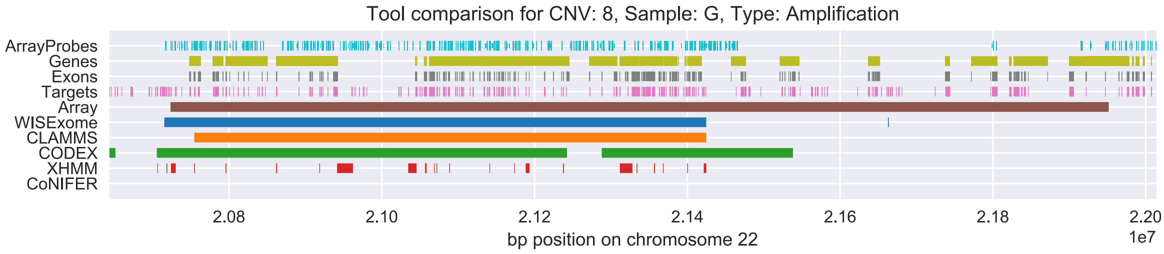

Tool comparison for CNV: 3, Sample: B, Type: Deletion
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
0
500000
1000000
1500000
2000000
2500000
bp position on chromosome 7

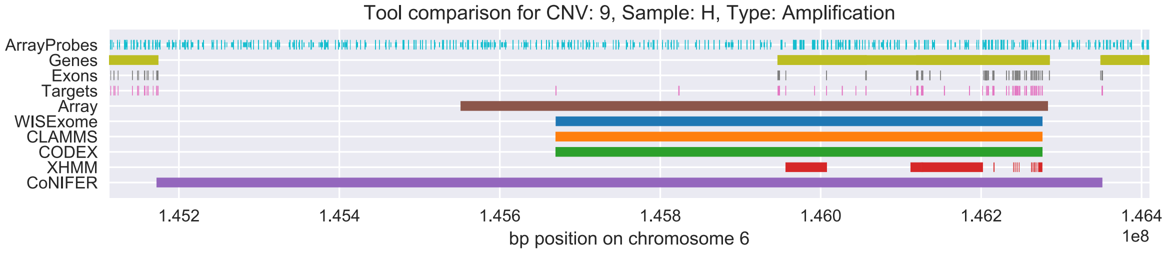

Tool comparison for CNV: 4, Sample: C, Type: Deletion
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
3.100
3.125
3.150
3.175
3.200
3.225
3.250
3.275
3.300
1e7
bp position on chromosome 15

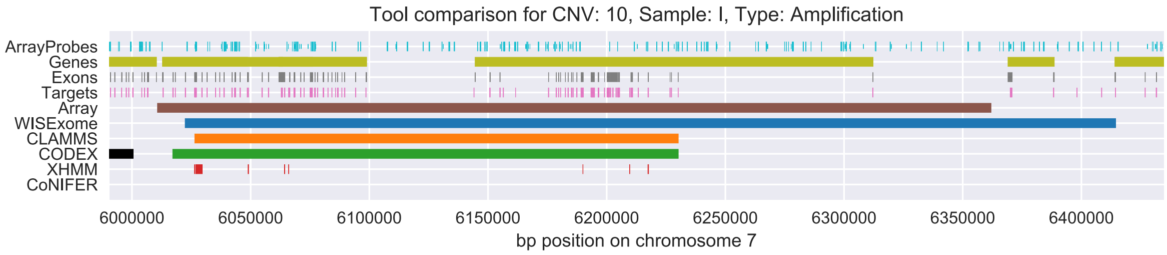

Tool comparison for CNV: 5, Sample: D, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
0.875
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1e7
bp position on chromosome 8

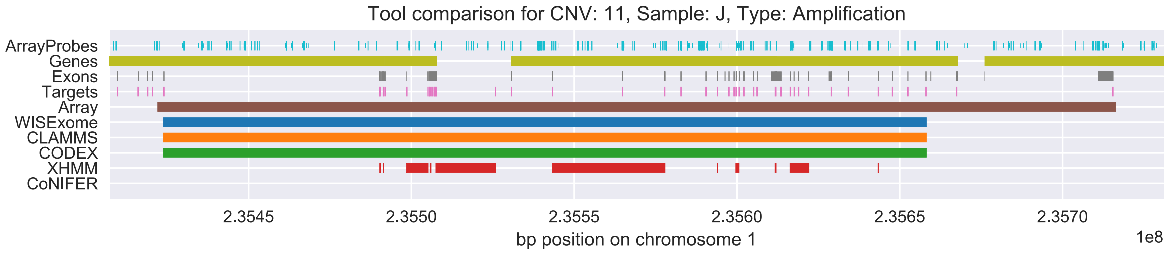

Tool comparison for CNV: 6, Sample: E, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
6500000
6750000
7000000
7250000
7500000
7750000
8000000
bp position on chromosome X

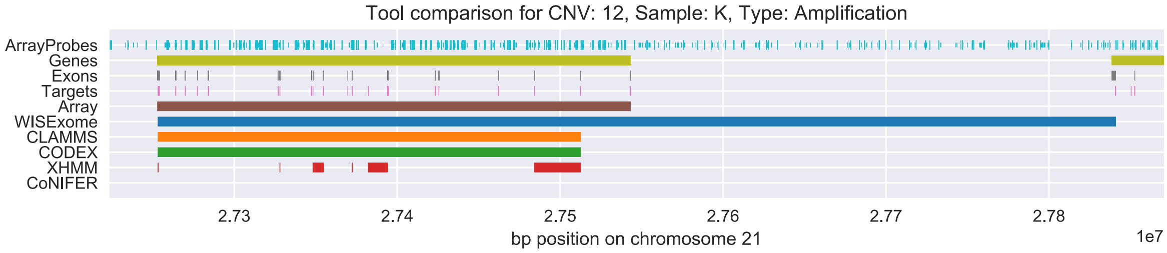

Tool comparison for CNV: 7, Sample: F, Type: Deletion
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
500000
750000
1000000
1250000
1500000
1750000
2000000
2250000
bp position on chromosome 20


Tool comparison for CNV: 8, Sample: G, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
2.08
2.10
2.12
2.14
2.16
2.18
2.20
1e7
bp position on chromosome 22

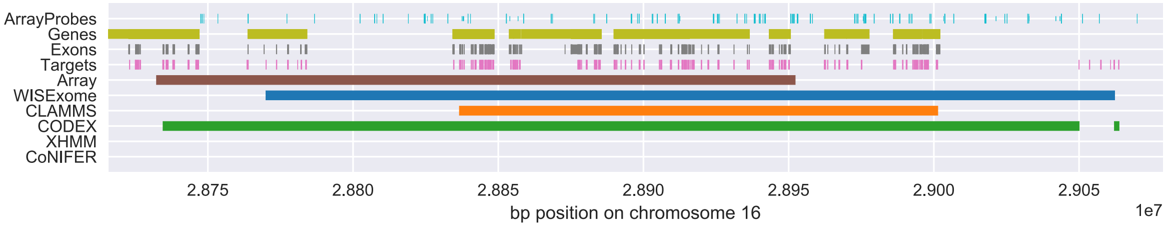

Tool comparison for CNV: 9, Sample: H, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
1.452
1.454
1.456
1.458
1.460
1.462
1.464
1e8
bp position on chromosome 6


Tool comparison for CNV: 10, Sample: I, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
6000000
6050000
6100000
6150000
6200000
6250000
6300000
6350000
6400000
bp position on chromosome 7

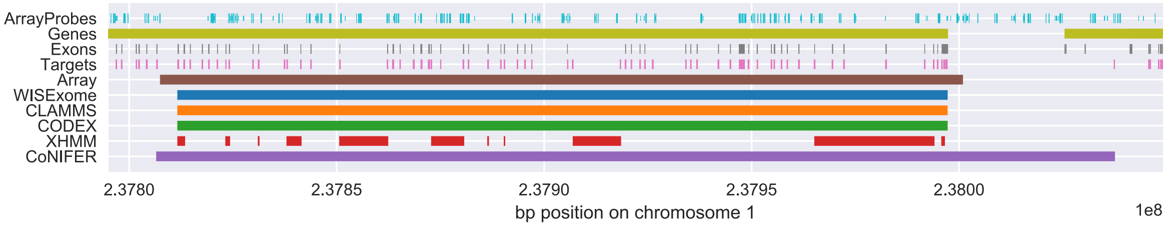

Tool comparison for CNV: 11, Sample: J, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
2.3545
2.3550
2.3555
2.3560
2.3565
2.3570
1e8
bp position on chromosome 1


Tool comparison for CNV: 12, Sample: K, Type: Amplification
ArrayProbes
Genes
Exons
Targets
Array
WISExome
CLAMMS
CODEX
XHMM
CoNIFER
2.73
2.74
2.75
2.76
2.77
2.78
1e7
bp position on chromosome 21

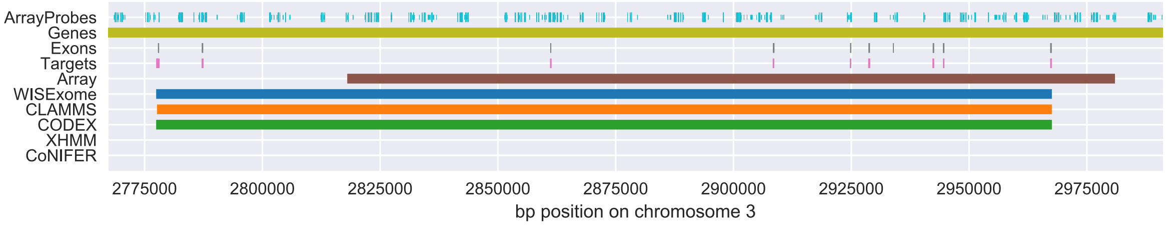
Tool comparison for CNV: 13, Sample: L, Type: Deletion



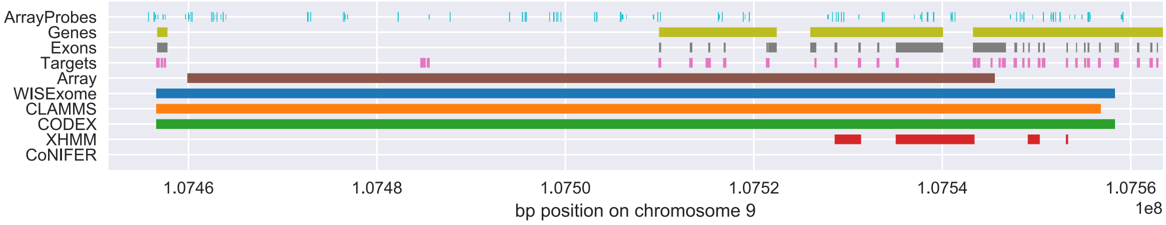
Tool comparison for CNV: 14, Sample: J, Type: Amplification



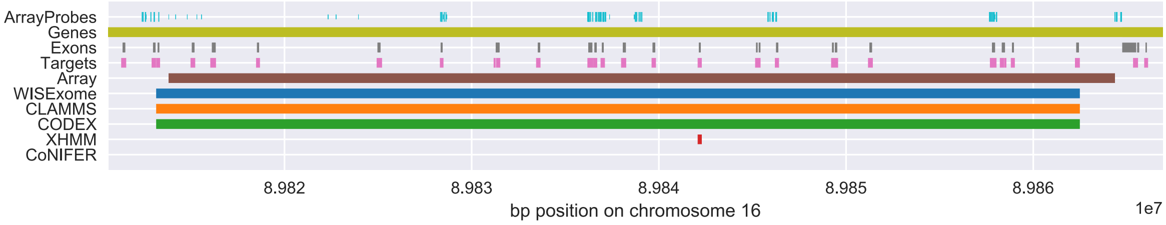
Tool comparison for CNV: 15, Sample: M, Type: Deletion



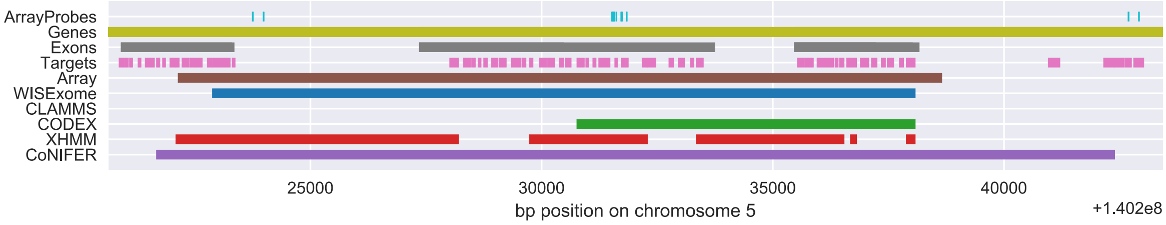
Tool comparison for CNV: 16, Sample: N, Type: Deletion



Tool comparison for CNV: 17, Sample: I, Type: Deletion



Tool comparison for CNV: 18, Sample: O, Type: Deletion



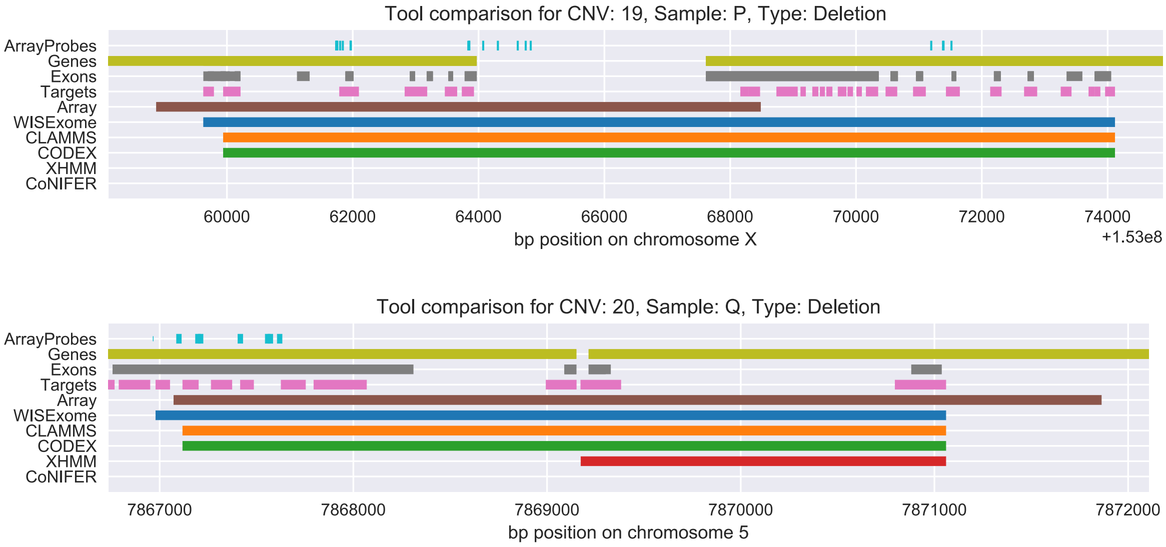


Figure S4: Detected CNV segments for all tools for all CNVs reported by clinical geneticist using array analysis. CNV numbers and sample identifiers match those in Supplementary Table S1. The region is annotated by the array probes (cyan), genes (citron), exons (grey) and target regions (pink).

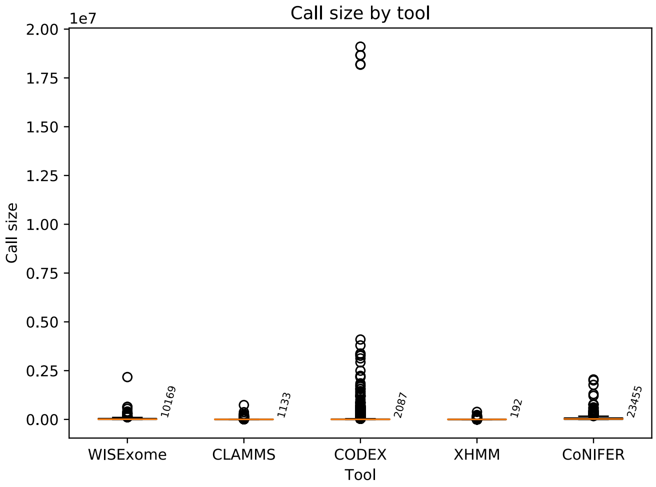


Figure S5: The same boxplot as shown in Figure 3a in the manuscript, except no cropping was performed. Numbers inside the figure annotate the actual median for every tool.

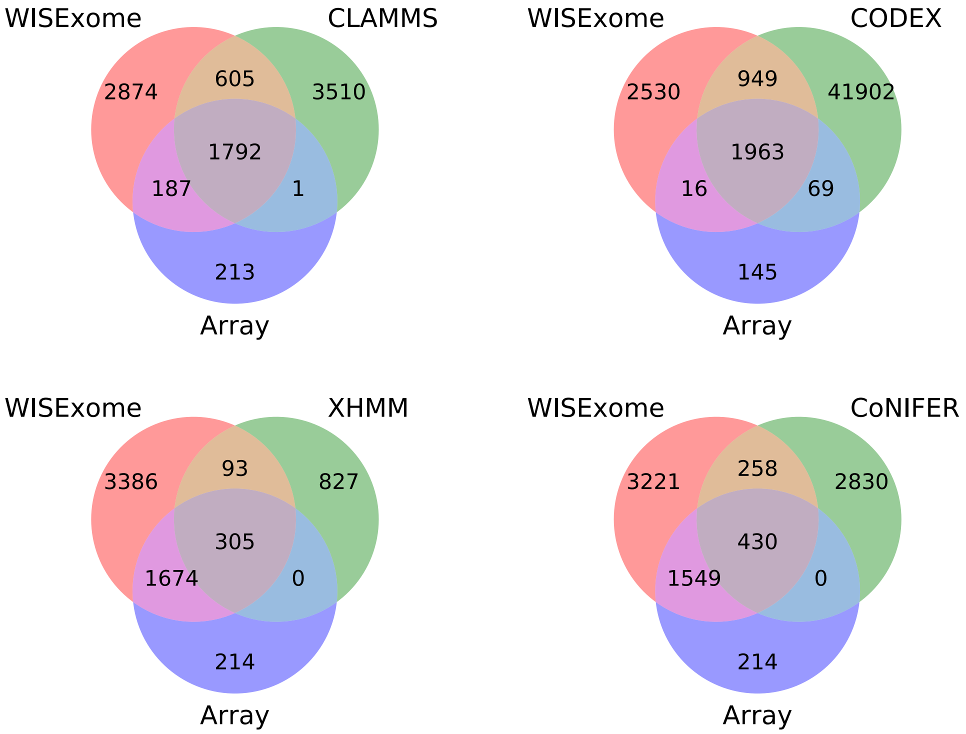


Figure S6: Overlap of exons affected by calls by WISExome, array analysis and (a) CLAMMS, (b) CODEX, (c) XHMM, and (d) CoNIFER.

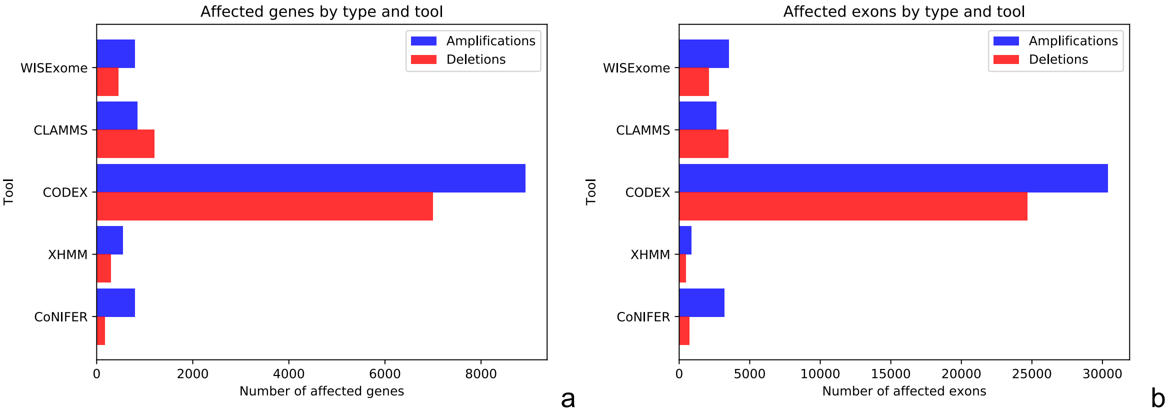


Figure S7: Number of genes and exons affected by amplifications and deletions per tool after thresholding. (a) Number of genes affected by calls per tool. (b) Number of exons affected by calls per tool.

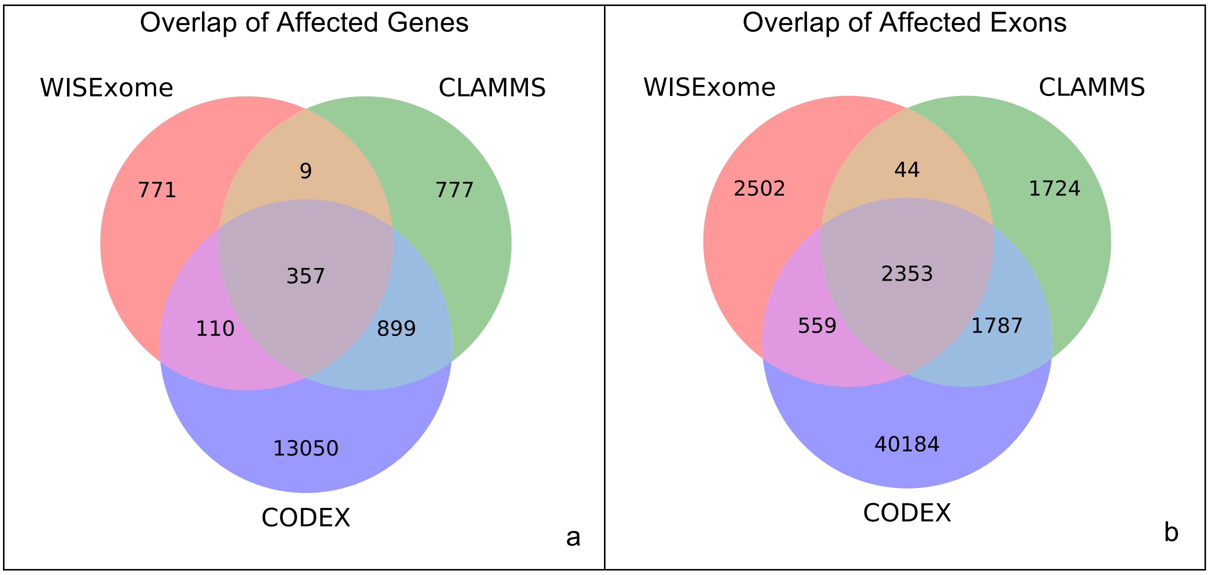


Figure S8: Overlap of (a) genes and (b) exons affected by calls by WISExome, CLAMMS and CODEX.

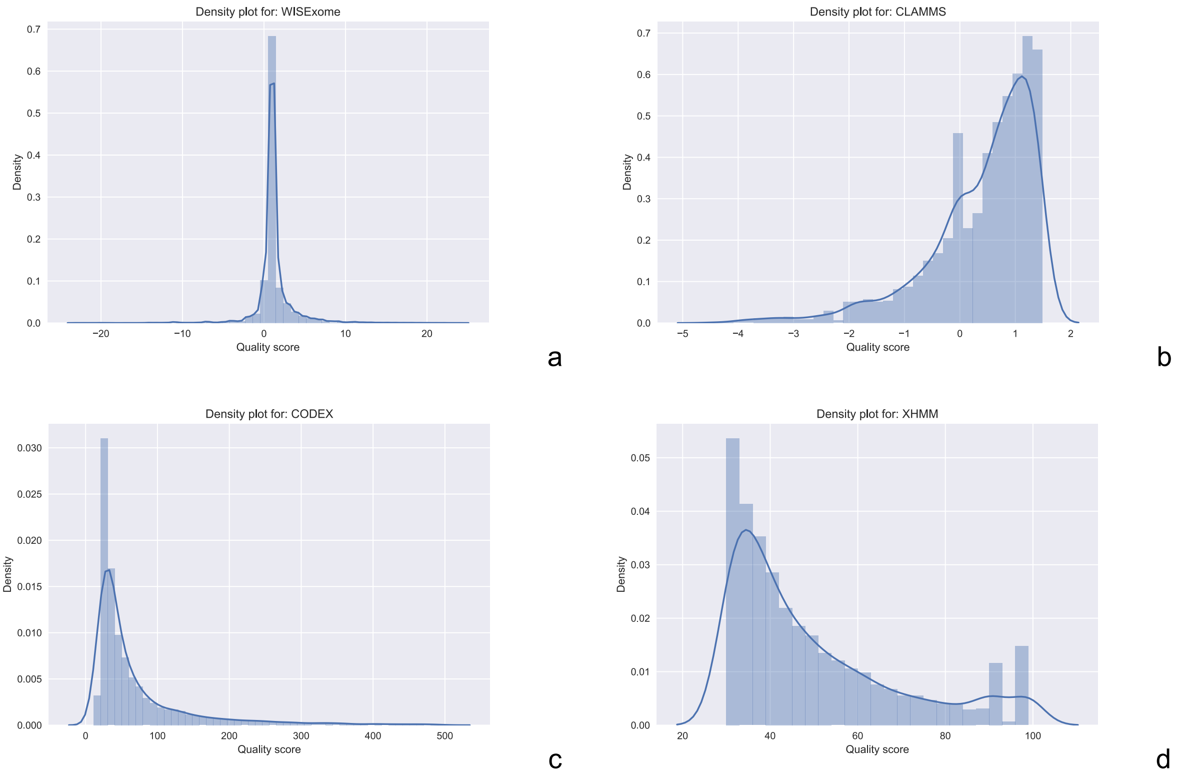


Figure S9: Density plots for the quality scores that are generated by the different CNV detection tools for the calls that they generate. CoNIFER is excluded as it does not provide a quality score per region. (a) WISExome's, (b) CLAMMS's, (c) CODEX's and (d) XHMM's quality score distributions. Every distribution is fitted with a kernel density estimate.

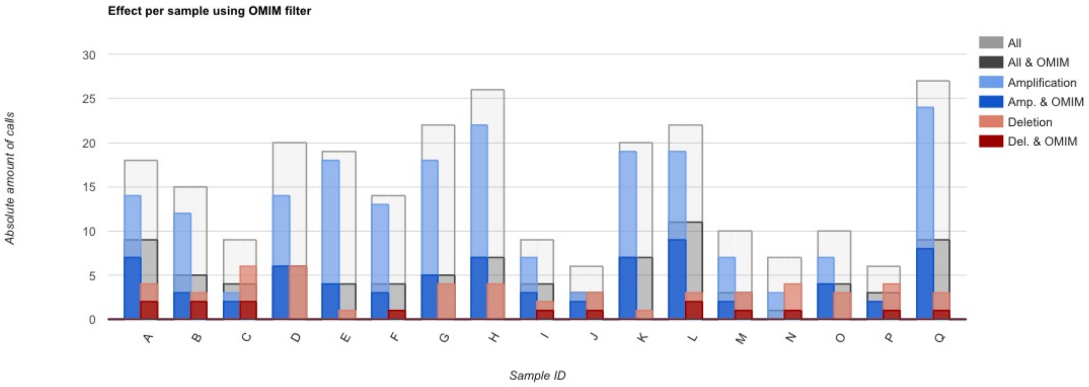


Figure S10: Effect on number of calls per sample when filtering on WISExome's quality score and OMIM phenotype key >= 3. Data is shown for all calls (gray), and for amplifications (blue) and deletions (red) separately. Light colors show the effect of applying only the quality filter, darker colors show the amounts of calls when overlap with at least one gene with an OMIM phenotype key of >= 3 is also required.

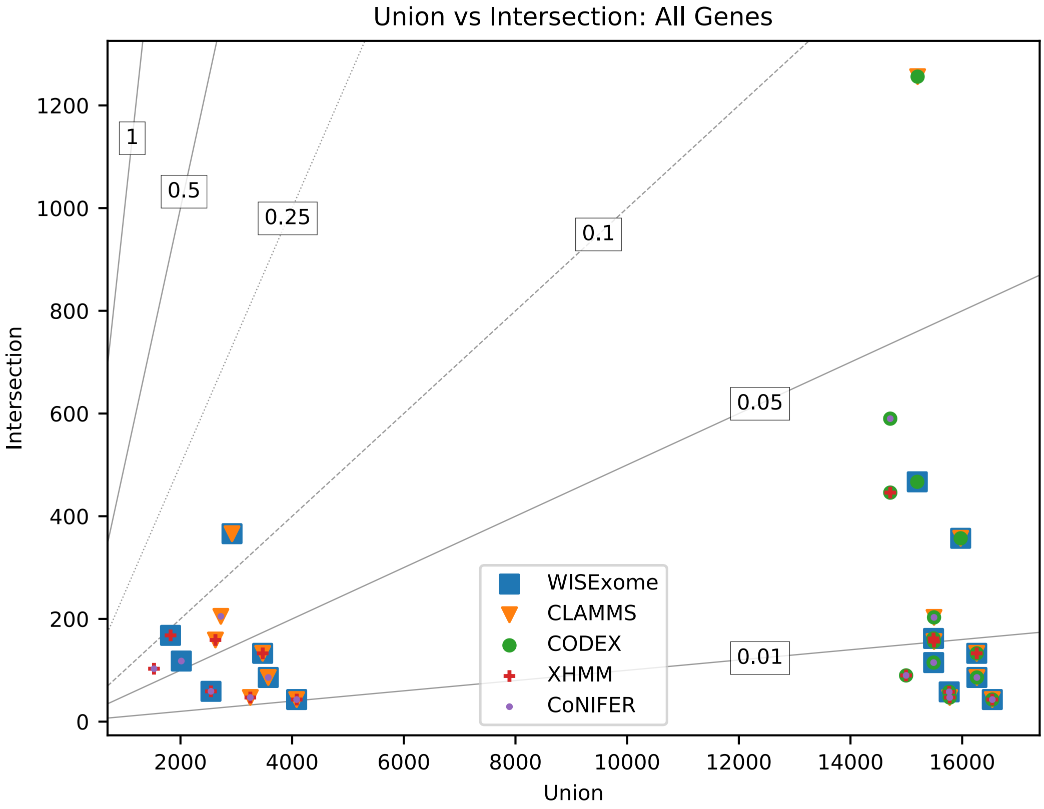


Figure S11: Plot showing the overlap in detected genes between any combination of tools. The figure is the equivalent of Figure 4b where overlap in exons is shown instead. For a detailed description of the figure we therefore refer to Figure 4b.

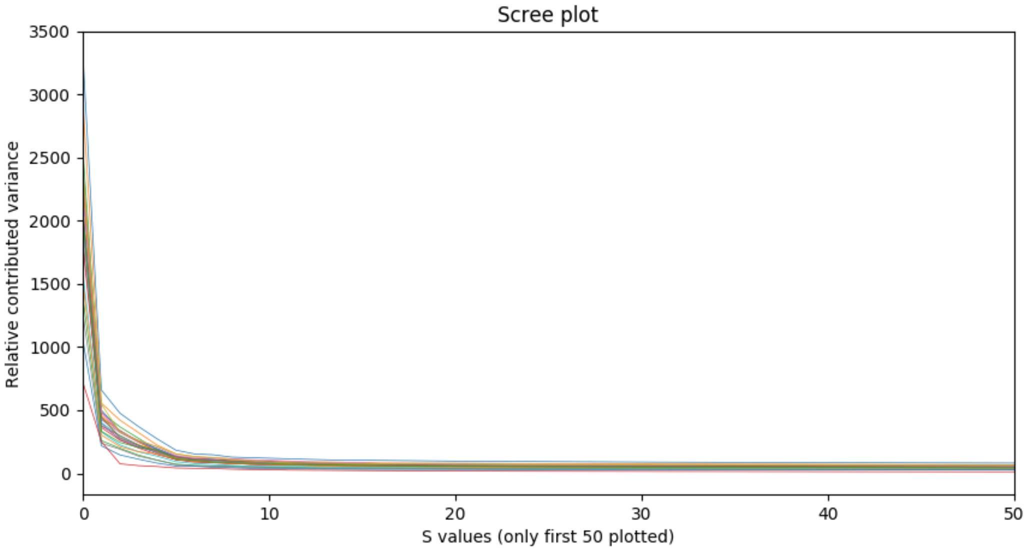


Figure S12: The scree plot used to determine the value of the SVD variable for CoNIFER.

SM1. Corrected significance threshold

Here we derive the used thresholds for significance.

First, we applied FWER using these settings:

α	0.05	Significance threshold
N	366795	Number of targets
w	8	Number of windows

Multiplying the probe count by the number of windows tested we obtained the total number of tests (T) done:

$$T = N \cdot w \tag{S1}$$

We determined the p-value required for this number of tests:

$$\frac{1}{T} \cdot \alpha \tag{S2}$$

Which equals to:

$$\frac{\alpha}{N \cdot w} \tag{S3}$$

Using the numbers for our settings we obtain:

$$\frac{0.05}{366795 \cdot 8} = 1.7\mathrm{e} - 8 \tag{S4}$$

Looking up the corresponding z-score for a two-tailed p-value of 1.7e-8 we found a corresponding z-score threshold of 5.64

SM2. Student's t-test with a pooled variance

This part is meant to explain the mathematical background to the segmentation algorithm we applied to fine-tune the borders of detected aberrations.

For this we use the following variables:

$X1$	samples on the segmented region (considered to be aberrated)
$X2$	samples outside the segmented region (not aberrated)
m_1	mean of samples of $X1$
s_1^2	variance of samples of $X1$
m_2	mean of samples of $X2$
s_2^2	variance of samples of $X2$

We use the following formula for a t-test with unequal sample sizes and equal variance:

$$t - test = \frac{(m_1 - m_2)}{s_p} \tag{S5}$$

where:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}} \tag{S6}$$

For $X2$ we sample values from the whole chromosome except for the candidate aberated regions. Consequently, n_2 is very large in comparison to n_1. This provides the rule:

$$n_2 \gg n_1 > 1 \tag{S7}$$

with this rule, we can approximate the $\sqrt{}$ term in the denominator of the t-test (equation S1), i.e.

(S8)

Now the t-test can be rewritten to:

$$t - test = \frac{\sqrt{n_1}}{s_p(m_1 - m_2)} \tag{S9}$$

Rewriting s_p:

$$s_p = \sqrt{\frac{(n_1 - 1)}{n_1 + n_2 - 1} s_1^2 + \frac{(n_2 - 1)}{n_1 + n_2 - 1} s_2^2} \tag{S10}$$

Again, using the rule stated in equation (S7), $n_2 \gg n_1$, $n_2 \gg 1$, this can be approximated as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)}{n_2} s_1^2 + \frac{(n_2 - 1)}{n_2} s_2^2} \tag{S11}$$

Using that $n_2 \gg 1$, the first term disappears. The ratio in the second term equals 1. From that it follows that the pooled variance becomes:

$$s_p = s_2 \tag{S12}$$

Since the non aberrated region on the chromosome is very large s_2 will be nearly constant when the segment $(X1)$ changes size. Therefore, when maximizing the t-test s_p can be considered constant:

$$t^* = \frac{\sqrt{n_1}}{s_p}(m_1 - m_2) \rightarrow \sqrt{n_1}(m_1 - m_2) \qquad \text{(S13)}$$

Our implementation maximizes the difference between the mean of the segment and all other, not aberrated, target regions on the same chromosome, multiplied by the square root of the length of the segment.

From the deduction shown here, it follows our segmentation equals optimization between two sets using the t-test.

SM3. Settings used for other tools

XHMM: Samples supplied to XHMM (downloaded from GitHub @ 18 June 2015) were the same BAM files as prepared for WISExome testing. XHMM was run according to the tutorial. The quality filter was set to Q_SOME >= 60, as was suggested in the tutorials. XHMM counts the read coverage per exon and employs Principal Component Analysis (PCA) to remove most of the technical variations over a set of samples. XHMM takes several heuristics into account such as exome-wide CNV rates, length distributions and the distance between target regions, all of these were set to default values. XHMM could not distinguish between training and test samples, thus allowing test samples to influence their own results and other samples through both the PCA and sample comparisons.

CoNIFER: CoNIFER (version 0.2.2; Released 9/17/2012) was provided all samples involved in this project. Settings were kept at defaults, with the --svd variable set to 4. Seeing the results were missing many of the known positives, we also tried running with SVD values of 2 up to and including 6 to ensure we did not misinterpret the scree plot (Supplementary Figure S12), but these alternative settings did not change the results significantly. Just like XHMM, CoNIFER could not distinguish between training and test samples, thus allowing test samples to influence their own results and other samples through both the SVD and sample comparisons.

CODEX: Codex (GitHub, commit 3d40ac9 @ April 7 2017) was run according to the vignette and tutorials supplied. As we have a clear group of affected individuals and controls, we used the normalize2 function which allowed us to specify the healthy control group. CODEX appears to make many calls allowing the end user to set a threshold. As their paper states, CODEX found significantly more calls than XHMM (2-fold) and CoNIFER (10-fold) in their tests, our finding appears in line with the original authors statements. As we were unable to find a suggestion for this threshold in the manual and tutorials, we decided not to apply one to influence the results as little as possible.

CLAMMS: We ran CLAMMS (GitHub, commit 3e19892 @ April 10 2017) with all default settings and suggested values. Based on the readme, we applied Q_EXACT >= 0 as a threshold. Results shown here were created using all reference data as reference set. We did not manually pick a reference set per sample (as suggested in the readme), as this would make results too dependent on user interaction. We tried the automated k-tree that was also described, but using the settings shown in the examples results became far worse, likely due to using a too small reference set per test sample.

For tools that provided a scoring per region, we plotted a distribution of these values in Supplementary Figure S9.