Supplemental Materials to "Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2"

Romina D'Aurizio [1], Tommaso Pippucci [2], Lorenzo Tattini [3], Betti Giusti [3], Marco Pellegrini[1], Alberto Magi [3]

[1]LISM, Institute of Informatics and Telematics and Institute of Clinical Physiology, National Research Council, Pisa [2]Medical Genetics Unit, Sant'Orsola Malpighi Polyclinic, Bologna [3]Department of Computer Science, University of Pisa, Pisa and [4]Department of Experimental and Clinical Medicine, University of Florence, Florence

August 8, 2016

Supplemental Methods and Data

Here we extended the RC approach (Magi *et al.*, 2010) to all the genome for the identification of CNVs from WES data. We proved, indeed, that around 30% on average of reads produced by WES experiments align outside the targeted regions. As for In-target regions, also for Off-target if the sequencing process is uniform then the number of reads mapping to each genomic region is expected to be proportional to the number of times the region appears in the DNA sample. Following this assumption, the copy number of any genomic region can be estimated by counting the number of reads aligned to non-overlapping and contiguous genomic windows of predefined size L. Even though the sequencing processes producing In- and Off-target reads are independent we can apply the RC approach also to Off-target regions. To this end we expanded the Exon Mean Read Count (EMRC) that we introduced in 2013 (Magi *et al.*, 2013) defining the Window Mean Read Count (WMRC) which account for both In-target exons and Off-target windows of fixed size.

1 GC-content and Mappability normalisation of WMRC

In our previous work (Magi *et al.*, 2013) we demonstrated that EMRC data are affected by three sources of biases: the total GC content, the genomic mappability and the exon size which we removed using the median normalisation approach introduced by Yoon in 2009 (Yoon *et al.*, 2009) for GC content and extended by us for mappability and exon size biases. Here we proved that WMRC data in Off-target windows are affected by similar GC content and mappability biases that we corrected according to the following formula:

$$\overline{WMRC}_w = WMRC_w \frac{m}{m_X} \tag{1}$$

where $WMRC_w$ is the number of reads aligned to a genomic region of length W. W varies according to the size of each targeted exon or, in case of Off-target, it corresponds to the selected fixed size of non-overlapping windows in which the intergenic chromosome is divided; m_X is the median WMRC of all the windows that have the same X value (where X stands for GC content, mappability score and, only for In-target, exon size) as the i-th window, and m is the overall median of all the windows.

2 Dataset description

2.1 WES survey dataset analysis

The WES survey dataset is made of 30 different sequencing experiments performed using three enrichment kits (Agilent SureSelect Human All Exon 50Mb, NimbleGen SeqCap EZ Exome v2.0 44Mb, Illumina TruSeq Exome Enrichment 62Mb). The raw sequences of ERR039174, SRR309292, SRR309293 and SRR309291 were downloaded from NCBI Sequence Read Archive while 26 WES were sequenced by our group using Illumina HiSeq2000 system and producing 100PE reads libraries. Details are summarised in Supp. Table 1. All reads were aligned to the human reference genome (hg19) using BWA short read aligner 0.7.5-r404 (Li and Durbin, 2010) and duplicates were filtered out with MarkDuplicates of Picard 1.92 package (`http://picard.sourceforge.net`). Read pairs with identical external coordinates could have been introduced by the PCR amplification step. MarkDuplicates retains only the pair with the highest mapping quality. Subsequently IndelRealigner module of GATK v2.5-2 (DePristo *et al.*, 2011) was used for local realignment around Indels. Reads with mapping quality MQ 10 were removed by using the SAMtools package 0.1.19-44428cd (Li *et al.*, 2009). Finally, we download the exact genomic coordinates of each target kit (Agilent SureSelect Human All Exon 50Mb, NimbleGen SeqCap EZ Exome v2.0 44Mb, Illumina TruSeq Exome Enrichment 62Mb) from manufacturer websites and we counted the number of reads mapping in In-Target, Flanking regions of 200bp or Off-Target by means of SAMtools flagstat.

2.2 1000 Genomes Project dataset

The accuracy of our method for predicting the absolute number of DNA copies of genomic regions from exome data in a population study was evaluated by analysing 8 WES from healthy individuals sequenced by the 1000 Genomes Project (1000 Genomes Project Consortium *et al.*, 2010) (see Supp. Table 2 for details). Same individuals had been previously genotyped by McCarroll (McCarroll *et al.*, 2008) and Conrad (Conrad *et al.*, 2010) using array-based technologies. The bam files were retrieved from the official 1KG ftp site(`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/`), and were processed, sorted and filtered (discarding MQ 10) with SAMtools. Last, PCR duplicates were removed using MarkDuplicates of Picard 1.92 package (`http://picard.sourceforge.net`). We calculated the normalized WMRC for In-target and Off-target as previously described, then the log2-ratio of normalised WMRC between each test and the control (NA10847) was analyzed by the SLM (Magi et al., 2010) (with $\omega = 0.1$, $\eta = 10^{-5}$)). The resulting valued were compared to the log2-copy number ratio of regions inferred by McCarroll and Conrad studies. Correlation with McCarroll calls are shown in Figure 3 of the main test, instead the results from the comparison with Conrad are in Supp. Figure 3.

Moreover, we analysed further 45 samples sequenced by the 1KG Project. The WES bam files were downloaded from the ftp site(`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/`). The original bam file were then sorted and filtered as previously described from duplicates. This dataset consists of exomes that were sequenced in 4 different centres (BCM - Baylor College of Medicine; BI - Broad Institute; BGI - Beijing Genomics Institute; WUGC - Washington University Genome Center) using different exome enrichment kits. The corresponding bed file of the designed targeted regions were retrieved from the resources listed at the 1KG website (`http://www.1000genomes.org/category/pull-down/`). The bam file were processed as previous described for the small dataset of 8 sample. The log2-ratio of normalised WMRC was analysed using the SLM segmentation algorithm (Magi *et al.*, 2010) followed by the FastCallSeq algorithm (Benelli *et al.*, 2010) (setting cellularity parameter c = 1). Same exome data was also analysed with XHMM (Fromer *et al.*, 2012), CoNIFER (Krumm *et al.*, 2012), CODEX (Jiang *et al.*, 2015) and CopywriteR Kuilman *et al.* (2015) and results were compared and evaluated using the collections of all genomic CNVs genotyped by HapMap and 1KG Project for the same 45 individuals. Precision and recall, as defined in the main text, were calculated and shown in Figure 4.

2.3 Urothelial bladder cancer samples

We tested our computational pipeline on a dataset including 14 bladder cancer tissues and 14 blood samples from same patients with different tumor stages and grades. Whole-exome sequencing data were downloaded from the SRA repository, study SRP029936. They are part of the study published by Balvas-Martinez and her colleagues in 2013 (Balbas-Martinez *et al.*, 2013). Agilent SureSelect Human All Exon

plus v3 50Mb or v4 51Mb were used for library preparation and enrichment (see Suppl. Table 4 for details). Sequencing was performed on HiSeq2000 producing 75bp paired-end reads. WES reads were aligned as described for the WES survey dataset (Sec. 2.1). Bed files of targeted regions were acquired directly from Agilent website. WES data was analysed with EXCAVATOR2 and CopywriteR (see next sections for details).

Same samples were also assayed using Illumina OmniExpress v1.0 SNP-array. For them, the log2ratio (red/green) values of single SNP for each sample were calculated. The ratios between each pair of tumour and control samples were then normalised and analysed using the SLM segmentation algorithm (Magi et al., 2010) (with $\omega = 0.1$, $\eta = 10^{-5}$). Finally each segment was classified by using the FastCall calling procedure (Benelli *et al.*, 2010), setting cellularity parameter c = 1.

Results by using 50K bp as size for the windows in off target regions were selected because the difference between In- and Off-Target SNR ratios were the lowest on average per sample. We compared the segmented signals from EXCAVATOR2 (HSLMResults_*.txt) and CopywriteR (from segment.RData) to segmented SLM results of SNP-arrays. madDiff function of CRAN matrixStat v0.14.2 package was employed for MAD value calculation.

3 Running parameters other tools

The dataset of 45 health individuals was analysed using XHMM v1.0, CONIFER v0.2.2, CODEX v0.99.6 and CopywriteR v1.99.4. We run XHMM with default parameters following authors' instructions in the tutorial (`http://atgu.mgh.harvard.edu/xhmm/tutorial.shtml`). Also for CONIFER we followed tutorial instructions (`http://conifer.sourceforge.net/tutorial.html`) setting the number of svd components removed to 2, based on the inflection point or the plateau of the scree plot generated. For CODEX, since, as suggested by the authors, it works better with similar experiments. We split the 45 WES dataset in 4 different subset according to sequencing library length (e.g. 76bp, 90bp,100bp and 101bp). Finally, we used CopywriteR with bin size = 50000bp and default parameters. Produced segmented profiles of each sample were extracted and transformed in copy number.
For the somatic dataset, CopywriteR was run with default parameters, leaving 20000bp of bin size.

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–73.

Balbas-Martinez, C., Sagrera, A., Carrillo-de Santa-Pau, E., Earl, J., Marquez, M., Vazquez, M., Lapi, E., Castro-Giner, F., Beltran, S., Bayes, M., Carrato, A., Cigudosa, J. C., Dominguez, O., Gut, M., Herranz, J., Juanpere, N., Kogevinas, M., Langa, X., Lopez-Knowles, E., Lorente, J. A., Lloreta, J., Pisano, D. G., Richart, L., Rico, D., Salgado, R. N., Tardon, A., Chanock, S., Heath, S., Valencia, A., Losada, A., Gut, I., Malats, N., and Real, F. X. (2013). Recurrent inactivation of stag2 in bladder cancer is not associated with aneuploidy. *Nature Genet.*, **45**(12), 1464–U221.

Benelli, M., Marseglia, G., Nannetti, G., Paravidino, R., Zara, F., Bricarelli, F. D., Torricelli, F., and Magi, A. (2010). A very fast and accurate method for calling aberrations in array-cgh data. *Biostatistics*, **11**(3), 515–8.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Wellcome Trust Case Control Consortium, Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, **464**(7289), 704–12.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, **43**(5), 491–8.

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P., and Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet*, **91**(4), 597–607.

Jiang, Y., Oldridge, D. A., Diskin, S. J., and Zhang, N. R. (2015). Codex: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res*, **43**(6), e39.

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., NHLBI Exome Sequencing Project, Quinlan, A. R., Nickerson, D. A., and Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res*, **22**(8), 1525–32.

Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., Nevedomskaya, E., Xu, G., de Ruiter, J., Lolkema, M. P., Ylstra, B., Jonkers, J., Rottenberg, S., Wessels, L. F., Adams, D. J., Peeper, D. S., and Krijgsman, O. (2015). Copywriter: Dna copy number detection from off-target sequence data. *Genome Biol*, **16**, 49.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, **26**(5), 589–95.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16), 2078–9.

Magi, A., Benelli, M., Marseglia, G., Nannetti, G., Scordo, M. R., and Torricelli, F. (2010). A shifting level model algorithm that identifies aberrations in array-cgh data. *Biostatistics*, **11**(2), 265–80.

Magi, A., Tattini, L., Cifola, I., D'Aurizio, R., Benelli, M., Mangano, E., Battaglia, C., Bonora, E., Kurg, A., Seri, M., Magini, P., Giusti, B., Romeo, G., Pippucci, T., De Bellis, G., Abbate, R., and Gensini, G. F. (2013). Excavator: detecting copy number variants from whole-exome sequencing data. *Genome Biol*, **14**(10), R120.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. (2008). Integrated detection and population-genetic analysis of snps and copy number variation. *Nat Genet*, **40**(10), 1166–74.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, **19**(9), 1586–92.

Sample ID	WES kit	total reads	total reads (no dup)	dup (%dup)	Mapped In Target (%)	Flanking (%)	Off target (%)
ERR039174	NimGen SC	88347712	66898351	21449361 (24.28)	47478507 (70.97)	9252149 (13.83)	10167695 (15.20)
SRR309292	NimGen SC	184983780	165113236	19870544 (10.74)	124714393 (75.53)	15880036 (9.62)	24518807 (14.85)
SC1	NimGen SC	52261424	47051646	5209778 (9.97)	18170372 (38.62)	2691579 (5.72)	26189695 (55.66)
SC2	NimGen SC	66880930	64410978	2469952 (3.69)	44665665 (69.34)	8628131 (13.40)	11117182 (17.26)
SC3	NimGen SC	64261764	61674574	2587190 (4.03)	43174166 (67.03)	7757938 (12.58)	10742470 (17.42)
SC4	NimGen SC	97504234	94011659	3492575(3.58)	33713495 (54.66)	8901725 (9.47)	51396439 (54.67)
SC5	NimGen SC	54596812	52705854	3492575 (3.58)	33426555 (63.42)	7305126 (13.86)	11974173 (22.72)
SC6	NimGen SC	53105762	51332720	1773042 (3.34)	33352701 (64.97)	7191335 (14.01)	10788684 (21.02)
SC7	NimGen SC	52428800	50766702	1662098 (3.17)	33226087 (65.45)	7233795 (14.25)	10306820 (20.30)
SC8	NimGen SC	51585664	49573371	2012293 (3.9)	33015415 (66.60)	7556660 (15.24)	9001296 (18.16)
SRR309293	Illum TS	112885944	99846786	13039158 (11.55)	48697849 (48.77)	7786469 (7.80)	43362468 (43.43)
TS1	Illum TS	98066566	67599801	30466765 (31.07)	41498127 (61.39)	10066857 (14.89)	16034817 (23.72)
TS2	Illum TS	149040972	107487177	41553795 (27.88)	66251120 (61.64)	16472076 (15.32)	24763981 (23.04)
TS3	Illum TS	233371152	149634324	83736828 (35.88)	89456899 (59.78)	24944108 (16.67)	35233317 (23.55)
TS4	Illum TS	204814304	135320695	69493609 (33.93)	81506931 (60.23)	21977943 (16.24)	31835821 (23.53)
TS5	Illum TS	15204305	14713126	491179(3.23)	9073611 (61.67)	1572408 (10.69)	3718292 (25.27)
TS6	Illum TS	119472318	89972393	29499925 (24.69)	53982638 (60.00)	13807865 (15.35)	22181890 (24.65)
TS7	Illum TS	144394810	109084718	35310092 (24.45)	66439634 (60.91)	15164919 (13.90)	27480165 (25.19)
TS8	Illum TS	113879490	80482975	33396515 (29.33)	50288688 (62.48)	10909671 (13.56)	19284616 (23.96)
TS9	Illum TS	109680894	81855834	27825060 (25.37)	49949128 (61.02)	11145849 (13.62)	20760857 (25.36)
SRR309291	Agil SS	124112466	104169460	19943006 (16.07)	73377796 (70.44)	8182048 (7.85)	22609616 (21.70)
SS1	Agil SS	49506550	46343163	3163387 (6.39)	24407118 (52.67)	6716967 (14.49)	15219078 (32.84)
SS2	Agil SS	62748490	55762934	6985556 (11.13)	33388926 (59.88)	7796468 (13.98)	14577540 (26.14)
SS3	Agil SS	68381472	55404594	12976878 (18.98)	34239385 (61.80)	8162598 (14.73)	13002611 (23.47)
SS4	Agil SS	60646496	53777154	6869342 (11.33)	32663249 (60.74)	3698186 (6.88)	17415719 (32.38)
SS6	Agil SS	65577064	57517689	8059375 (12.29)	35942669 (62.49)	8180373 (14.22)	13394647 (23.29)
SS7	Agil SS	80597930	73125655	7472275 (9.27)	40299478 (55.11)	10399825 (14.22)	22426352 (30.67)
SS8	Agil SS	84954462	77293918	7660544 (9.02)	41725253 (53.98)	11857340 (15.34)	23711325 (30.68)
SS10	Agil SS	68112888	53492078	14620810 (21.47)	31496528 (58.88)	6723084 (12.57)	15272466 (28.55)
SS11	Agil SS	88791162	78225154	10566008 (11.9)	38577180 (49.32)	11857177 (15.16)	27790797 (35.53)

Supplemental Table 1: Detailed statistics about the samples used for the WES Survey Data set. Exomes produced using NimbleGen kits have on average 71.5 million reads ranging from 5.1 to 18.5 millions reads, TruSeq exomes are of 142 millions reads on average varying from 98 to 233 millions and SureSelect ones range from 49 to 124 millions reads, 71 millions on average. NimGen SeqCap replaces NimbleGen SeqCap EZ Exome v2.0 44Mb, Illum TS replaces Illumina TruSeq Exome Enrichment 62Mb and Agil SS replaces Agilent SureSelect Human All Exon 50Mb.

		Conrad et al.				McCarroll et al.			
Sample	Population	DUP	DEL	mean length DUP	mean length DEL	DUP	DEL	mean length DUP	mean length DEL
NA19131	YRI	365	438	9334	9082	120	121	52652	31830
NA19138	YRI	347	476	9262	13013	112	110	49533	40549
NA19152	YRI	357	405	9378	8182	110	124	53074	33511
NA19159	YRI	351	405	8961	9092	118	130	46345	34391
NA19200	YRI	344	468	9954	11348	111	128	47474	32185
NA19206	YRI	371	405	9167	8596	107	126	48527	34878
NA19223	YRI	359	490	9582	10304	107	128	47702	27347
NA10847	CEU	-	-	-	-	-	-	-	
Mean		356.2	441	9376.8	9945.2	112.1	123.8	49329.5	33527.2

Supplemental Table 2: Samples from 1000 Genomes Project with Conrad's and McCarroll's CNV characterisation. NA10847 was used as control.

Sample	Population	Project Center	Platform	Tot Exome Seq	% Targets≥20x
NA06985	CEU	BCM	ILLUMINA	6027283676	90
NA06986	CEU	BI	ILLUMINA	24104123900	93
NA07000	CEU	BCM	ILLUMINA	6452445499	91
NA07048	CEU	BI	ILLUMINA	16685037188	92
NA07051	CEU	BI	ILLUMINA	14837397464	86
NA07347	CEU	BI	ILLUMINA	18370193548	88
NA07357	CEU	BCM	ILLUMINA	5529433769	89
NA10851	CEU	BCM	ILLUMINA	6819007021	93
NA11829	CEU	BCM	ILLUMINA	6612402027	92
NA11830	CEU	BCM	ILLUMINA	6749788186	93
NA11831	CEU	BCM	ILLUMINA	5723298522	89
NA11832	CEU	BCM	ILLUMINA	6894731266	92
NA11881	CEU	BCM	ILLUMINA	7996466435	94
NA11918	CEU	BI	ILLUMINA	23573702952	92
NA11919	CEU	BI	ILLUMINA	18583826356	83
NA11920	CEU	BI	ILLUMINA	16208787444	94
NA11992	CEU	BCM	ILLUMINA	6222852602	90
NA11993	CEU	BCM	ILLUMINA	6757616191	NA
NA11994	CEU	BCM	ILLUMINA	7010501910	92
NA11995	CEU	BCM	ILLUMINA	3843488239	84
NA12003	CEU	BCM	ILLUMINA	3944293208	84
NA12005	CEU	BCM	ILLUMINA	6636259439	91
NA12043	CEU	BCM	ILLUMINA	8878610737	93
NA12044	CEU	BCM	ILLUMINA	6857270164	93
NA12045	CEU	BI	ILLUMINA	15487027832	86
NA12144	CEU	BCM	ILLUMINA	6587073651	93
NA12154	CEU	BCM	ILLUMINA	6494580477	92
NA12156	CEU	BCM	ILLUMINA	4215485884	82
NA12234	CEU	BCM	ILLUMINA	6852784956	87
NA12489	CEU	BCM	ILLUMINA	6330614047	91
NA12762	CEU	BCM	ILLUMINA	5128453366	88
NA12812	CEU	BCM	ILLUMINA	6615884406	89
NA12813	CEU	BCM	ILLUMINA	9309388564	92
NA12815	CEU	BCM	ILLUMINA	5770682066	89
NA12872	CEU	BCM	ILLUMINA	5433437511	88
NA12873	CEU	BCM	ILLUMINA	6420725136	90
NA12878	CEU	BI	ILLUMINA	17082242372	89
NA18486	YRI	BI	ILLUMINA	17668337348	94
NA18489	YRI	BI	ILLUMINA	14711774252	87
NA18498	YRI	BI	ILLUMINA	6882363008	73
NA18499	YRI	BI	ILLUMINA	13181571784	93
NA18501	YRI	BI	ILLUMINA	13369756144	93
NA18502	YRI	BCM	ILLUMINA	7423286890	89
NA18504	YRI	BI	ILLUMINA	16864396732	88
NA18505	YRI	BCM	ILLUMINA	9692055445	92
NA18507	YRI	BCM	ILLUMINA	5139715977	84
NA18508	YRI	BCM	ILLUMINA	10881456291	92
NA18510	YRI	BI	ILLUMINA	17725889472	90
NA18516	YRI	BI	ILLUMINA	12915915760	93
NA18519	YRI	BI	ILLUMINA	9932791964	83
NA18520	YRI	BI	ILLUMINA	11531938204	91
NA18522	YRI	BI	ILLUMINA	23265974468	93
NA18631	CHB	BI	ILLUMINA	17008911204	88
NA18634	CHB	BI	ILLUMINA	17716920986	88
NA18639	CHB	BCM	ILLUMINA	6607324757	89
NA18640	CHB	BCM	ILLUMINA	5932760503	91

Supplemental Table 3: Samples from 1000 Genomes Project...to be continued

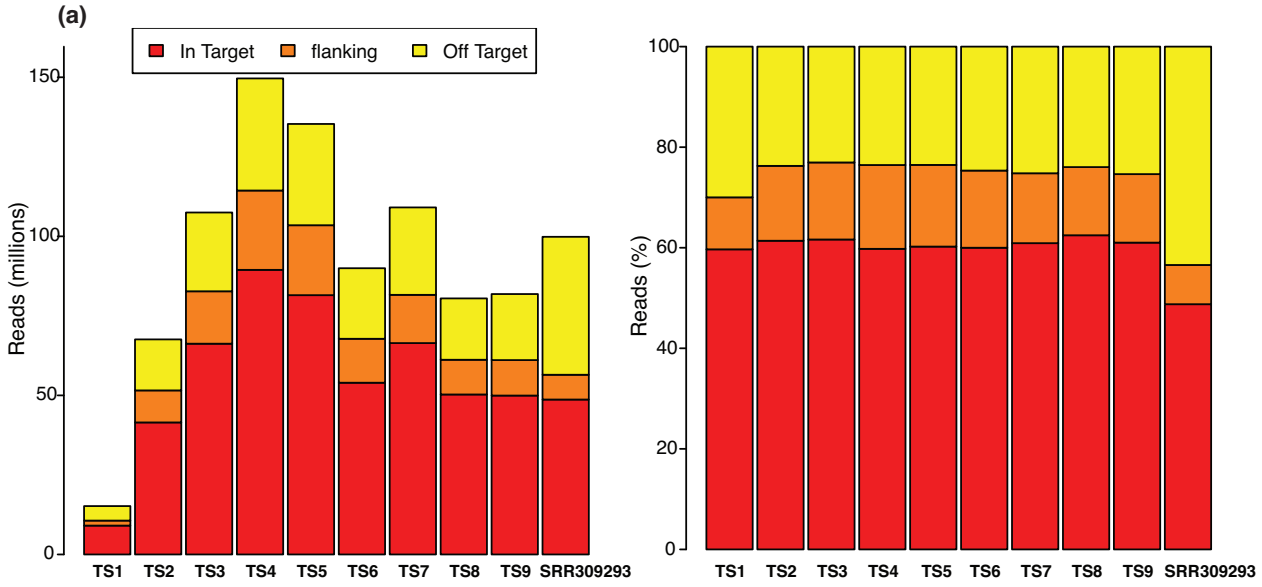
Sample	Population	Project Center	Platform	Tot Exome Seq	% Targets≥20x
NA18641	CHB	BCM	ILLUMINA	8638346685	92
NA18642	CHB	BCM	ILLUMINA	7414985801	90
NA18643	CHB	BCM	ILLUMINA	6767105545	91
NA18645	CHB	BCM	ILLUMINA	4918028451	86
NA18740	CHB	BCM	ILLUMINA	6884548042	91
NA18745	CHB	BCM	ILLUMINA	6026798775	90
NA18747	CHB	BCM	ILLUMINA	8257959677	93
NA18748	CHB	BCM	ILLUMINA	7223470308	90
NA18749	CHB	BCM	ILLUMINA	8422599171	92
NA18757	CHB	BCM	ILLUMINA	6337149858	90
NA18853	YRI	BI	ILLUMINA	16464420436	94
NA18856	YRI	BI	ILLUMINA	16425604272	87
NA18867	YRI	BI	ILLUMINA	21646713376	95
NA18868	YRI	BI	ILLUMINA	15737628332	90
NA18870	YRI	BI	ILLUMINA	13840647584	93
NA18871	YRI	BI	ILLUMINA	15020932828	87
NA18873	YRI	BI	ILLUMINA	23834535408	92
NA18874	YRI	BI	ILLUMINA	9492274604	83
NA18910	YRI	BI	ILLUMINA	14325967928	92
NA18912	YRI	BI	ILLUMINA	25547455784	93
NA18953	JPT	BI	ILLUMINA	20537617560	94
NA18956	JPT	BCM	ILLUMINA	9098365123	89
NA18959	JPT	BI	ILLUMINA	16045090892	93
NA18960	JPT	BI	ILLUMINA	19403095412	94
NA18961	JPT	BI	ILLUMINA	11694842684	92
NA18964	JPT	BI	ILLUMINA	18804225520	94
NA18965	JPT	BCM	ILLUMINA	6887697424	83
NA18978	JPT	BCM	ILLUMINA	7178040710	90
NA18980	JPT	BCM	ILLUMINA	10486917264	92
NA18997	JPT	BCM	ILLUMINA	6521651305	90
NA18998	JPT	BCM	ILLUMINA	8059427411	89
NA18999	JPT	BI	ILLUMINA	23249343692	93
NA19000	JPT	BI	ILLUMINA	17575878636	93
NA19003	JPT	BI	ILLUMINA	11462473596	92
NA19005	JPT	BCM	ILLUMINA	6193270611	89
NA19007	JPT	BI	ILLUMINA	16983811428	93
NA19012	JPT	BI	ILLUMINA	23514324076	93
NA19092	YRI	BI	ILLUMINA	10096466924	92
NA19093	YRI	BI	ILLUMINA	13066696796	93
NA19099	YRI	BCM	ILLUMINA	9409623085	92
NA19102	YRI	BI	ILLUMINA	14024421816	93
NA19116	YRI	BI	ILLUMINA	15582041360	92
NA19130	YRI	BI	ILLUMINA	23860580152	93
NA19172	YRI	BI	ILLUMINA	24450358924	92

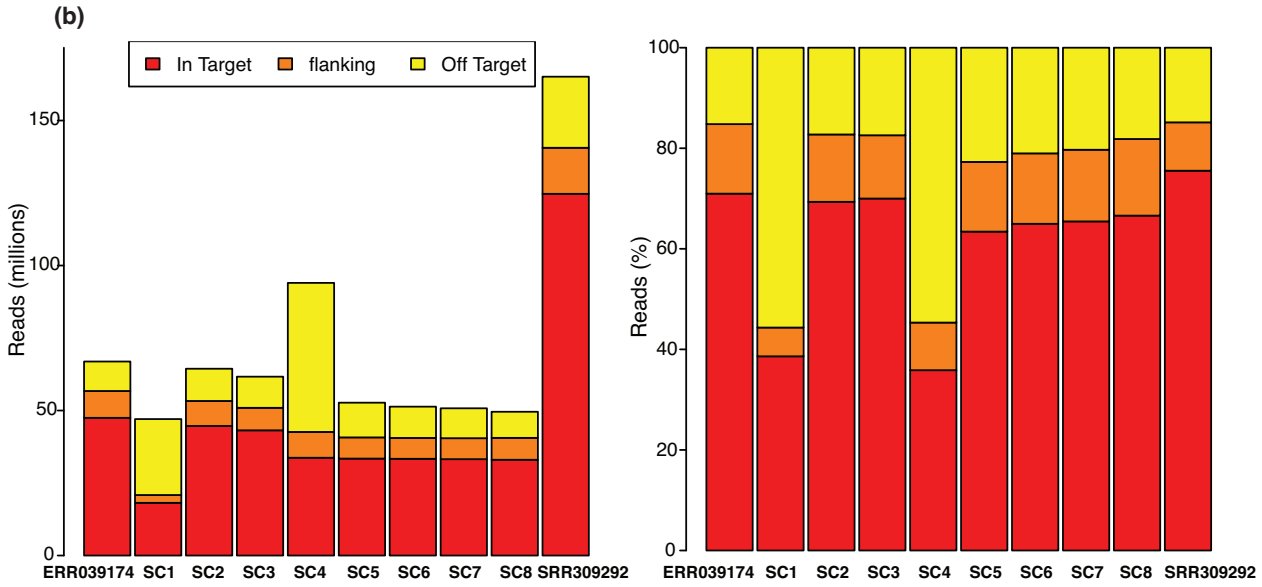
Supplemental Table 3: ...Samples from 1000 Genomes Project

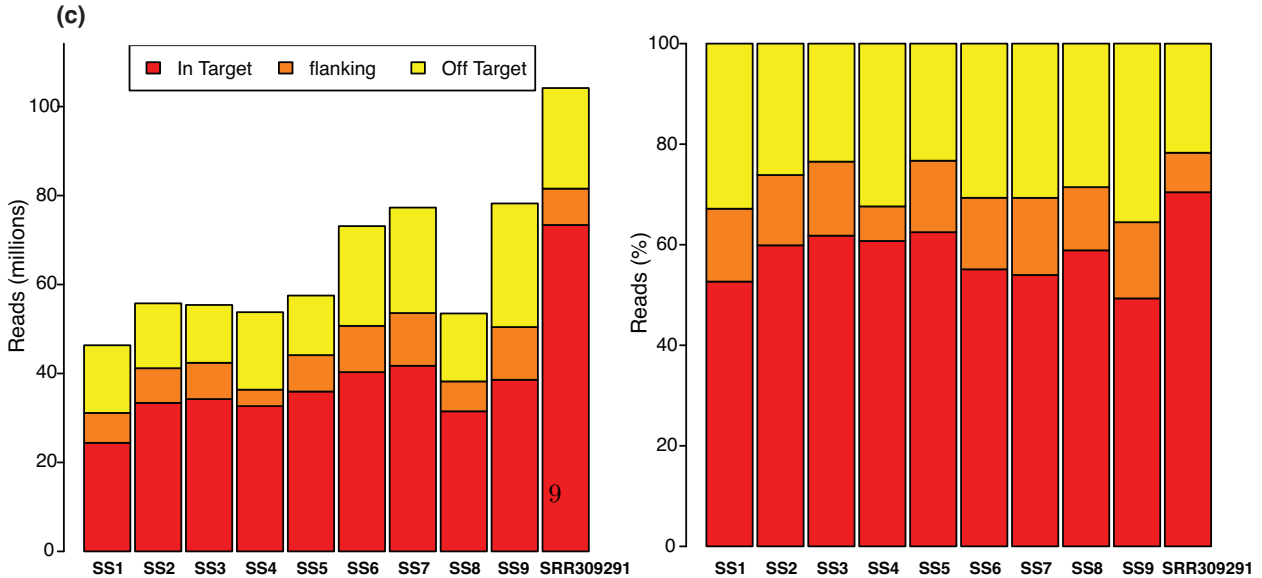
Run ID	Tissue Tipe	Target	Sample Name
SRR986735	Blood	v3	251-C
SRR986736	Bladder	v3	251-T
SRR986737	Blood	v3	114-C
SRR986738	Bladder	v3	114-T
SRR986739	Blood	v3	310-C
SRR986740	Bladder	v3	310-T
SRR986741	Blood	v3	116-C
SRR986742	Bladder	v3	116-T
SRR986743	Bladder	v3	331-C
SRR986744	Blood	v3	331-T
SRR986745	Blood	v3	413-C
SRR986746	Bladder	v3	413-T
SRR986751	Blood	v4	64-C
SRR986752	Bladder	v4	64-T
SRR986753	Blood	v4	179-C
SRR986754	Bladder	v4	179-T
SRR986755	Blood	v4	62-C
SRR986759	Bladder	v4	62-T
SRR986756	Blood	v4	188-C
SRR986760	Bladder	v4	188-T
SRR986757	Blood	v4	313-C
SRR986761	Bladder	v4	313-T
SRR986758	Blood	v4	343-C
SRR986762	Bladder	v4	343-T
SRR986763	Blood	v4	274-C
SRR986764	Bladder	v4	274-T
SRR986765	Blood	v4	451-C
SRR986766	Bladder	v4	451-T

Supplemental Table 4: Bladder Samples

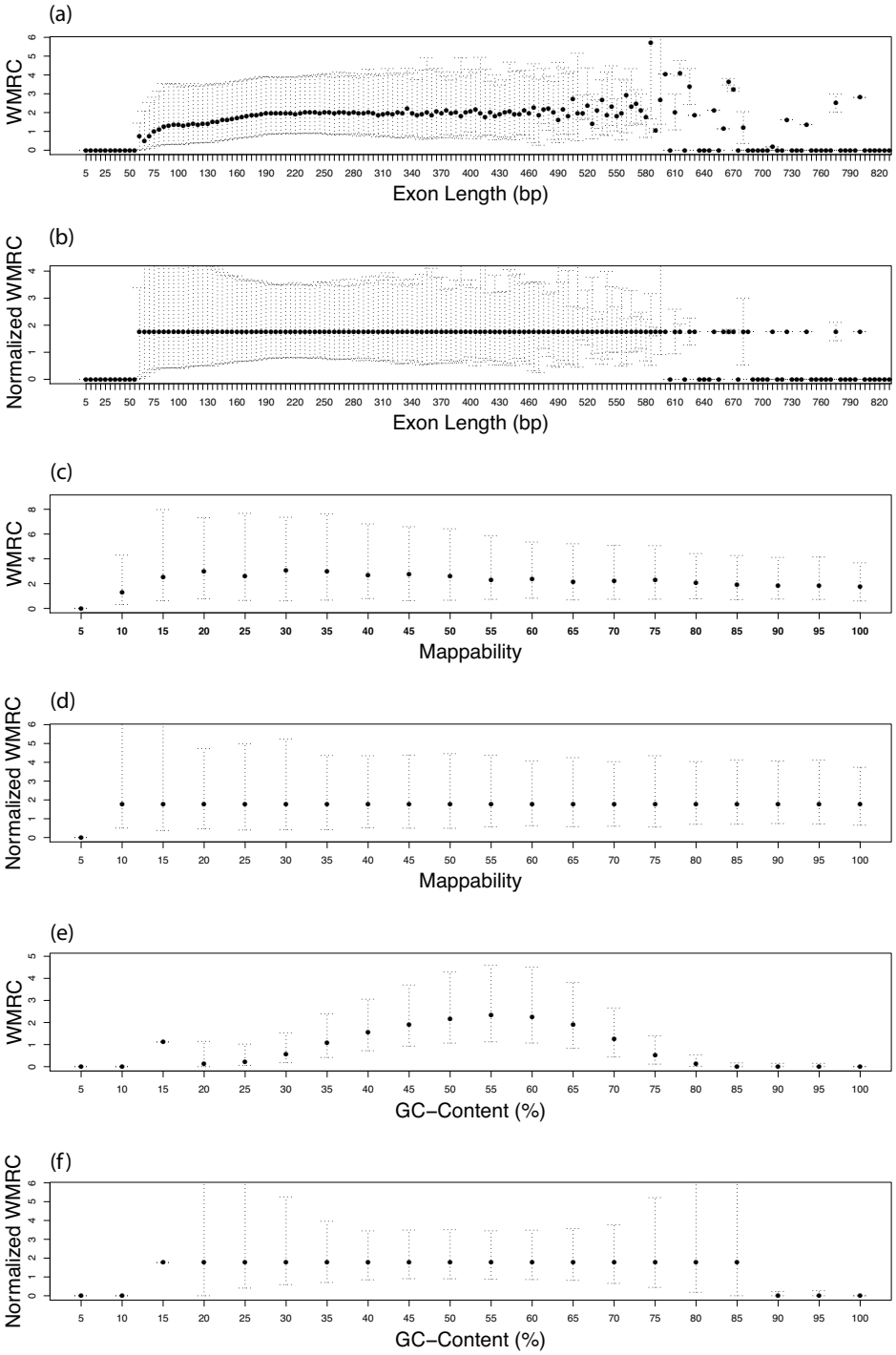
Supplemental Figure 1: Distribution of mapped reads for the 30 WES samples, split per enrichment kit : (a) Illumina TruSeq Exome Enrichment 62Mb (b) Agilent SureSelect Human All Exon 50Mb (c) NimbleGen SeqCap EZ Exome v2.0 44Mb. The absolute number (left) and relative number (right) of mapped reads to In-Target, Flanking and Off-Target regions are shown.



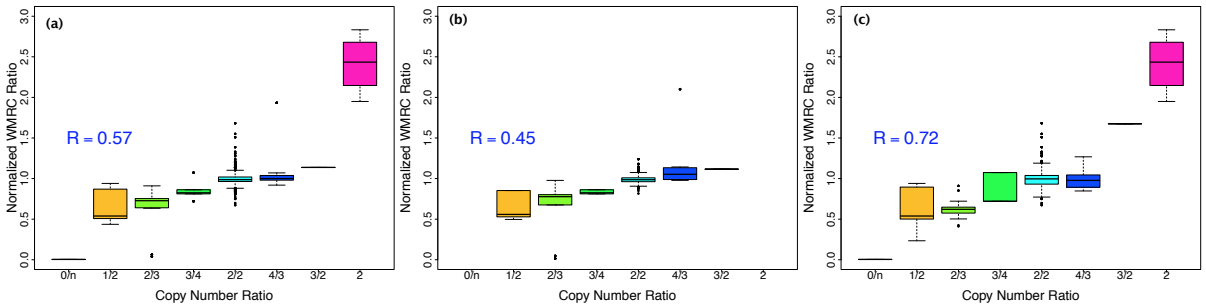




Supplemental Figure 2: Biases correction of WMRC for In-target regions. (a-b) pre and after exon-length normalisation, (c-d) pre and after mappability normalisation and (e-f) pre and after normalisation per local GC content



Supplemental Figure 3: Copy number correlation with Conrad calls. Boxplots summarize the capability of WMRC data to predict the exact number of DNA copies of a CNV region. Normalized WMRC ratio were calculated for eight samples using NA10847 as control and compared with copy number ratios from Conrad characterisation. R is the Pearson correlation coefficient. (a) all genomic regions, (b) In-Target regions and (c) Off-Target regions



Supplemental Figure 4: Deletions in bladder samples. The profiles of two regions containing potential deletions in Chr 3 and 6 of patient 64 are shown from SNParray (a,d), EXCAVATOR2 (b,e) and CopywriteR (c,f)