

The GENCODE exome – sequencing the complete human exome

Alison J. Coffey, Felix Kokocinski, Maria S. Calafato, Carol E. Scott, Priit Palta, Eleanor Drury, Christopher J. Joyce, Emily M. LeProust, Jen Harrow, Sarah Hunt, Anna-Elina Lehesjoki, Daniel J. Turner, Tim J. Hubbard, Aarno Palotie.

Supplementary content

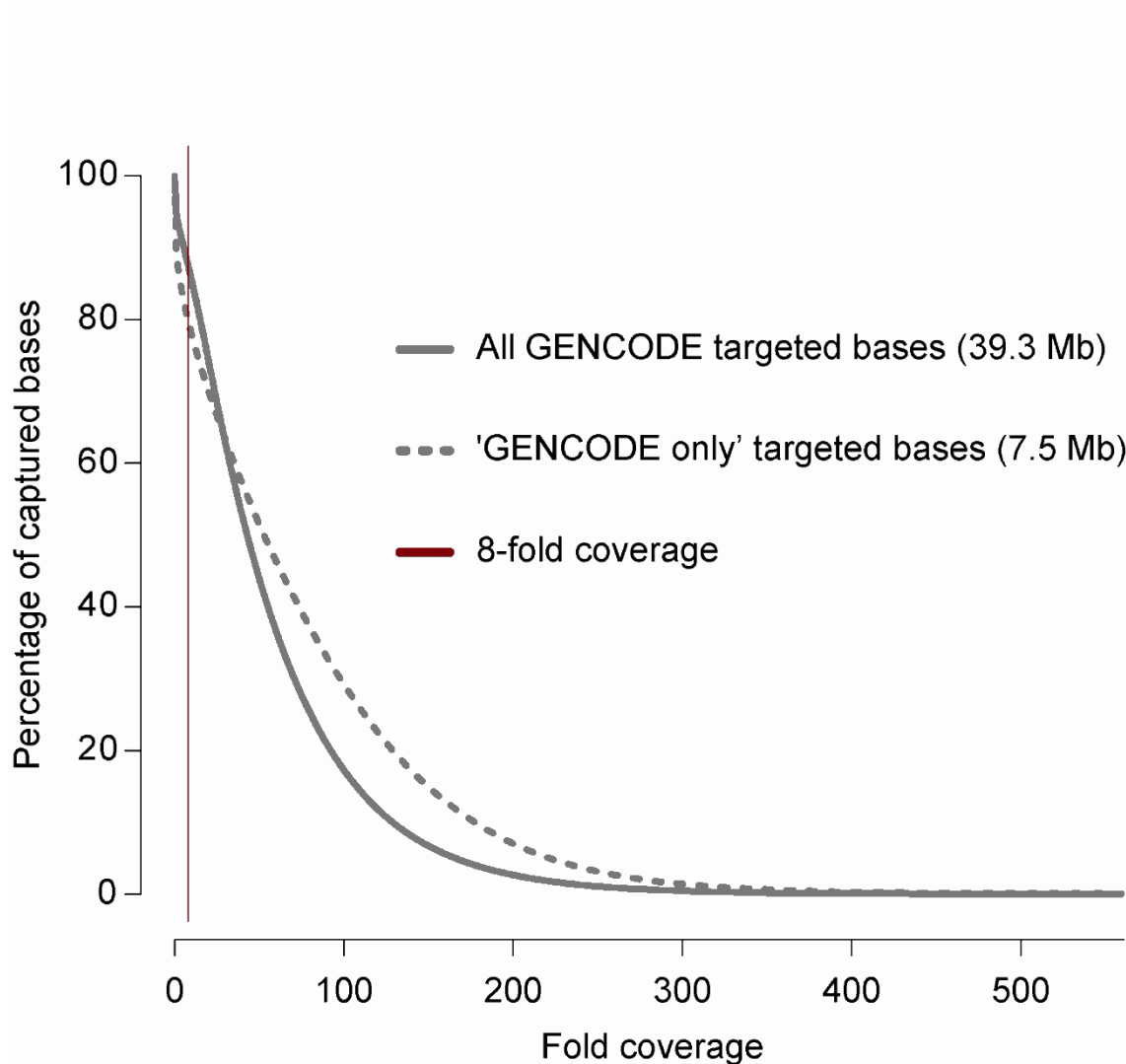
Figure 1	Cumulative coverage achieved by the GENCODE exome
Table 1	Comparison data between the three exome sets a. to external annotation b. to the design target
Table 2	Mapping characteristics (sequence yield and reads mapping back to target) from clinical and HapMap samples
Table 3	Repeat and low-complexity coverage
Table 4	Bait/probe covered CTR (Capture Target Regions) regions assessed using a uniqueness mask
Data 1	GENCODE exome design target
Data 2	GENCODE exome baits
Data 3	GENCODE exome exclusive regions
Data 4	GENCODE exome exclusive genes

Supplementary Figures

Supplementary Figure 1: Cumulative coverage achieved by the GENCODE exome.

Cumulative fold coverage plot for clinical samples captured with Agilent SureSelect Human All Exon Kit, the GENCODE exome; and the regions covered by the GENCODE exome only. The thin red vertical line indicates fold coverage of 8X, the minimum coverage required for variant calling.

Supplementary Figure 1



Supplementary Tables

Supplementary Table 1: Comparison data between the three exome sets

a. Comparison of external annotation (Gene Ontology, OMIM ids and HGNC names) based on Ensembl v57 for the three exome sets.

Exome set	CCDS exons	CCDS transcripts	RefSeq exons	RefSeq transcripts	GENCODE exons	GENCODE transcripts
NimbleGen Sequence Capture 2.1M Human Exome Array	82.80%	87.19%	79.47%	85.28%	75.29%	79.16%
Agilent SureSelect Human All Exon Kit	90.05%	93.16%	86.39%	90.91%	82.15%	84.67%
GENCODE exome	99.18%	99.63%	98.75%	99.19%	97.31%	96.51%
Additional content on GENCODE	9.12%		12.36%			

b. Coverage statistics of the design target by the three exome sets.

	gene number	transcript number	exon number
not covered	378	422	2856
covered by Agilent CCDS only	0	0	0
covered by Nimblegen CCDS only	10	12	318
covered by Nimblegen & Agilent CCDS only	0	0	0
covered by Nimblegen CCDS & Gencode exome only	365	873	4065
covered by Agilent CCDS & Gencode exome only	2202	5336	30708
covered by Gencode exome only	5594	9052	59600
covered by all 3	27828	65943	303483
<u>totals</u>	36853	82522	463778
Agilent CCDS	30030	71279	334191
Nimblegen CCDS	28203	66828	307866
Gencode exome	35989	81204	397856
<u>percents</u>			
Agilent CCDS	81.5%	86.4%	72.1%
Nimblegen CCDS	76.5%	81.0%	66.4%
Gencode exome	97.7%	98.4%	85.8%

Supplementary Table 2: Mapping characteristics (sequence yield and reads mapping back to target) from clinical and HapMap samples using GENCODE and Agilent CCDS exome captures.

Sample name	Sanger #1	Sanger #2	Sanger #3	Sanger #4	Sanger #5	Sanger #6	Sanger #7	NA12878	NA07000	NA19240	NA12878	NA07000	NA19240
Bait library	GENCODE							Agilent CCDS					
Lanes sequenced	3	3	3	3	3	3	3	4	3	3	3	4	4
Reads	105712968	115540748	107420522	101912732	129221672	205597102	203472750	249957386	216080714	175087650	131337304	174698654	206110412
Reads mapped all ($\geq q0$) ¹	101784111 (96.28%)	112290675 (97.19%)	104444274 (97.23%)	99877068 (98.00%)	126463612 (97.87%)	197922085 (96.27%)	196009151 (96.33%)	237675276 (95.09%)	212171108 (98.19%)	168827968 (96.42%)	123007651 (93.66%)	169580062 (97.07%)	198938967 (96.52%)
Reads mapped unique ($\geq q10$, duplicates removed) ²	77800441 (76.44%)	81310815 (72.41%)	74873012 (71.69%)	74979882 (75.07%)	97666019 (77.23%)	145520170 (73.52%)	143615416 (73.27%)	134894523 (56.76%)	122795772 (57.88%)	77220371 (45.74%)	91946948 (74.75%)	97248119 (57.35%)	117233076 (58.93%)
Unique reads mapped to CTR +/- 250bp ³	59625505 (76.64%)	69956576 (80.04%)	65714990 (87.77%)	63735274 (85.00%)	68908856 (70.56%)	86495628 (59.44%)	86910473 (60.52%)	124450772 (92.26)	113776953 (92.66%)	71794822 (92.97%)	83772118 (91.11%)	89399731 (91.93%)	99555617 (84.92%)
Unique reads mapped to GENCODE exome cluster regions (including 10bp flank) ³	40507262 (52.07%)	47280119 (58.15%)	46480228 (62.08%)	43032073 (57.39%)	45129213 (46.21%)	66486708 (67.96%)	66344444 (46.20%)	84566996 (62.69%)	81625999 (66.47%)	51683594 (66.93%)	56713564 (61.68%)	66088857 (67.96%)	69982605 (59.70%)
Mean depth GENCODE exome cluster regions (including 10bp flank)	46.38	54.08	53.45	49.83	51.02	76.96	76.61	95.37	92.37	58.81	65.08	75.92	80.14
Median depth GENCODE exome cluster regions (including 10 bp flank)	34	39	35	34	36	55	55	63	62	36	43	47	57
GENCODE exome cluster region bases $\geq 8x$ ⁴	32914133 (83.66%)	33102756 (84.14%)	32201108 (81.85%)	32536951 (82.70%)	32220983 (81.90%)	35330571 (89.80%)	34963040 (88.87%)	32415365 (82.39%)	32910800 (83.65%)	31502723 (80.07%)	28721876 (73.01%)	28903556 (73.47%)	29429202 (74.80%)

¹ Percentage of "Reads" in brackets

² Percentage of "Reads mapped all" in brackets

³ Percentage of "Reads mapped unique" in brackets

⁴ Percentage of GENCODE exome design target (39.3 Mb) in brackets

Supplementary Table 3: Assessment of repeat and low-complexity coverage of the three exome sets.

Repeats and low-complexity regions identified with RepeatMasker (parameters: *-nolow -species homo -s*), Dust and TRF using Ensembl 53 data.

Exome Set	total bp	bp with repeats	ratio
Nimblegen-CCDS	34,108,810	884,080	38.6
Agilent-CCDS	37,640,396	799,357	47.1
GENCODE exome	47,933,967	1,303,879	36.8

Supplementary Table 4: Bait/probe covered CTR (Capture Target Regions) regions assessed using a uniqueness mask developed by Heng Li for the 1000 Genomes project
(ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/README_hs36_uniqueness_mask)

Exome Supplier	CTR (Capture Target Regions) count	Total target size (bp)	Type 0 bases	Type 1 bases	Type 2 bases	Type 3 bases
Agilent	165,637 (316,000 baits)	37,640,396	0 (0.00%)	1,570,956 (4.17%)	838,459 (2.23%)	35,230,981 (93.60%)
Nimblegen	176,159 (197,218 baits)	34,108,807	17688 (0.05%) Y (X/Y masked)	1,236,875 (3.63%)	745,619 (2.19%)	32,108,625 (94.14%)
GENCODE exome	206,275 (406,539 baits)	47,933,967	0 (0.00%)	2,811,712 (5.87%)	1,177,904 (2.46%)	43,944,351 (91.68%)

Sequenceability/uniqueness measures used in the method:

- Type 0 : all 35mers covering this site cannot be mapped back due to "N"s in the reference
- Type 3 : $\geq 35 \times 0.5$ reads 2-away unique
- Type 2 : if not 3, $\geq 35 \times 0.5$ reads 1-away unique
- Type 1 : otherwise ($>35 \times 0.5$ reads are exact repeats)

Supplementary Data

(as separate files available at <ftp.sanger.ac.uk/pub/gencode/exome>)

Data 1: GENCODE exome design target

Genomic coordinates (NCBI36) of CDS regions of protein-coding genes and miRNAs (ECRs) from Ensembl and Havana selected as the design targets for the GENCODE exome in GTF 2.2 format. Annotated with Ensembl 53 as the source database.

Data 2: GENCODE exome baits

Genomic coordinates (NCBI36) of the baits of the GENCODE exome based on the design target in GTF 2.2 format. Annotated with Ensembl 58 as the latest information available.

Data 3: GENCODE exome exclusive regions

Genomic coordinates (NCBI36) of the bait regions present on the GENCODE exome, but missing from the Agilent and Nimblegen products in GTF 2.2 format; annotated with Ensembl 58 as the latest information available.

Data 4: GENCODE exome exclusive genes

List of genes present on the GENCODE exome, but missing from the Agilent and Nimblegen products in GTF 2.2 format using NCBI36.

The list contains genes originating from the Ensembl 53 database as well as genes from a snapshot of the Havana manual annotation database (Feb. 2009). The former were annotated with data from their source database, the latter from the VEGA database where possible. Genes from the PAR regions are listed with their X and Y locations. To avoid double-counting the numbers describing the GENCODE-only content in the text were based on Ensembl genes only, with PAR genes listed only once.