Supplementary Data

eXNVerify: coverage analysis for long and short read sequencing data in clinical context

1 Test data source

We utilized Extensive Sequence Dataset of Gold-Standard Samples for Benchmarking and Development from https://console.cloud.google.com/storage/browser/brain-genomics-public/research/sequencing/grch38/bam

2 Per-base BED file preparation

For example, HG003 sample from PacBio Hifi (BAM and BAI files) is downloaded and following mosdepth execution is performed to obtain per-base BED file:

```
docker run –it –v ~/data/pacbio_hifi/:/input \
quay.io/biocontainers/mosdepth:0.2.4––he527e40_0 \
mosdepth –t 32 ––fast–mode \
input/HG003.pacbio–hifi.21x.haplotag.grch38.bam \
input/HG003.pacbio–hifi.21x.haplotag.grch38.bam
```

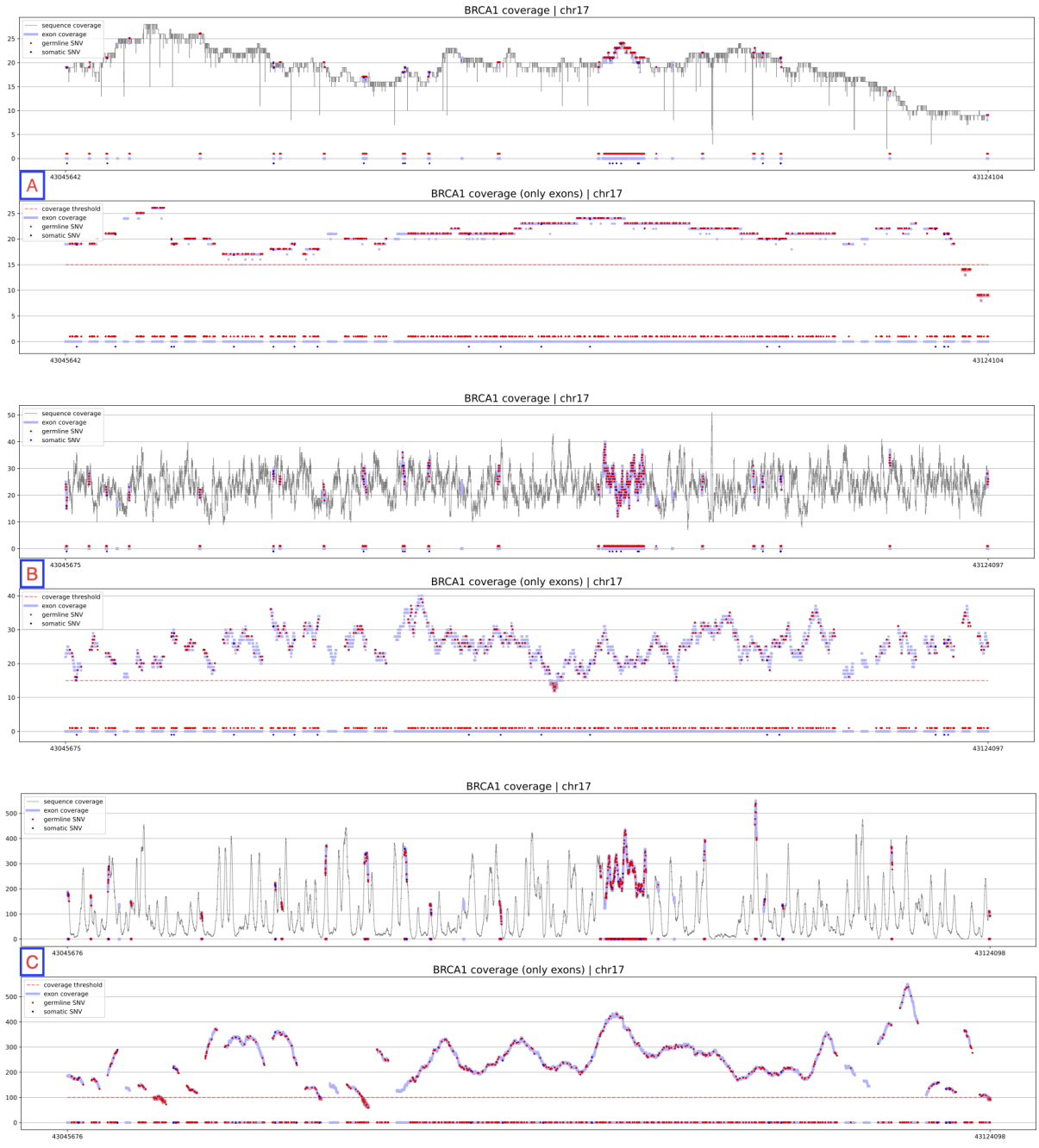
Then, per-base-bed.gz file is decompressed:

```
gzip –d HG003.pacbio–hifi.21x.haplotag.grch38.bam.per–base.bed.gz
```

Analogous per-base BED file generation should be done for all later samples.

3 Example of coverage analysis for BRCA1 gene and three considered sequence technologies

```
docker run –it ––rm –v ~/data/pacbio_hifi/:/input \
–v ~/data/pacbio_hifi/:/output –v ~/data/refs/:/refs \
porebskis/exnverify:1.0 ./geneCoverage.py \
input/HG003.pacbio–hifi.21x.haplotag.grch38.bam.per–base.bed \
refs/Exome_Reference_refined.bed refs/SNV_patho_germline.txt \
refs/SNV_patho_somatic.txt \
15 BRCA1

docker run –it ––rm –v ~/data/hiseqx_wgs_pcr_free/:/input \
–v ~/data/hiseqx_wgs_pcr_free/:/output –v ~/data/refs/:/refs \
porebskis/exnverify:1.0 ./geneCoverage.py \
input/HG003.hiseqx.pcr–free.20x.dedup.grch38.bam.per–base.bed \
refs/Exome_Reference_refined.bed refs/SNV_patho_germline.txt \
refs/SNV_patho_somatic.txt \
15 BRCA1

docker run –it ––rm –v ~/data/novaseq_wes_agilent/:/input \
–v ~/data/novaseq_wes_agilent/:/output –v ~/data/refs/:/refs \
porebskis/exnverify:1.0 ./geneCoverage.py \
input/HG003.novaseq.wes–agilent.100x.dedup.grch38.per–base.bed \
refs/Exome_Reference_refined.bed refs/SNV_patho_germline.txt \
refs/SNV_patho_somatic.txt \
100 BRCA1
```



Supplementary Figure 1: BRCA1 coverage for samples: A – PacBio Long Read, B – Illumina WGS, C – Illumina Exome Agilent

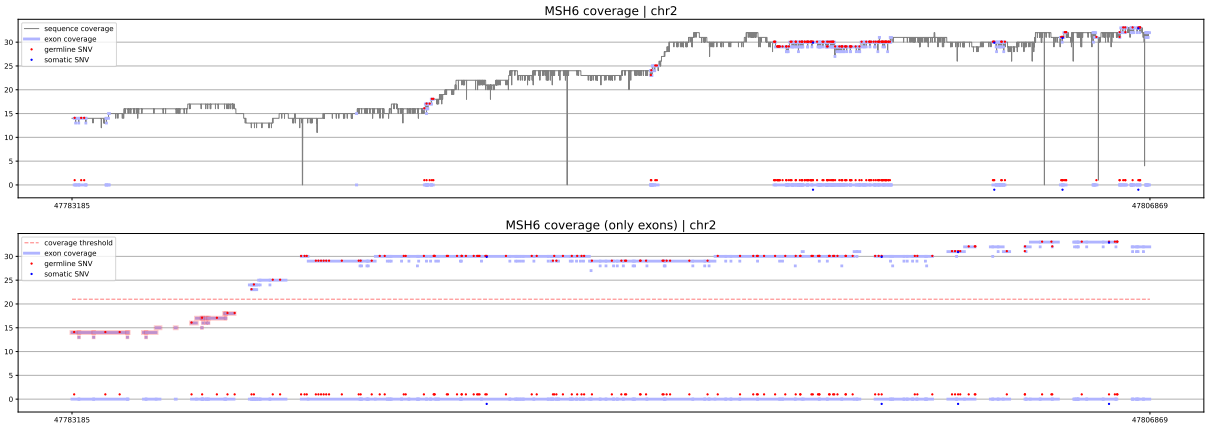
4 Example of coverage analysis for genes highly relevant for medical genetics and cancer genomics

```
docker run –it ––rm –v ~/data/pacbio_hifi/:/input \
–v ~/data/pacbio_hifi/:/output –v ~/data/refs/:/refs \
porebskis/exnverify:1.0 ./geneCoverage.py \
input/HG003.pacbio–hifi.21x.haplotag.grch38.bam.per–base.bed \
refs/Exome_Reference_refined.bed refs/SNV_patho_germline.txt \
refs/SNV_patho_somatic.txt \
21 MSH6 TP53 ABCA4 PDGFRB LEMD2 CFTR HTT DMD
```

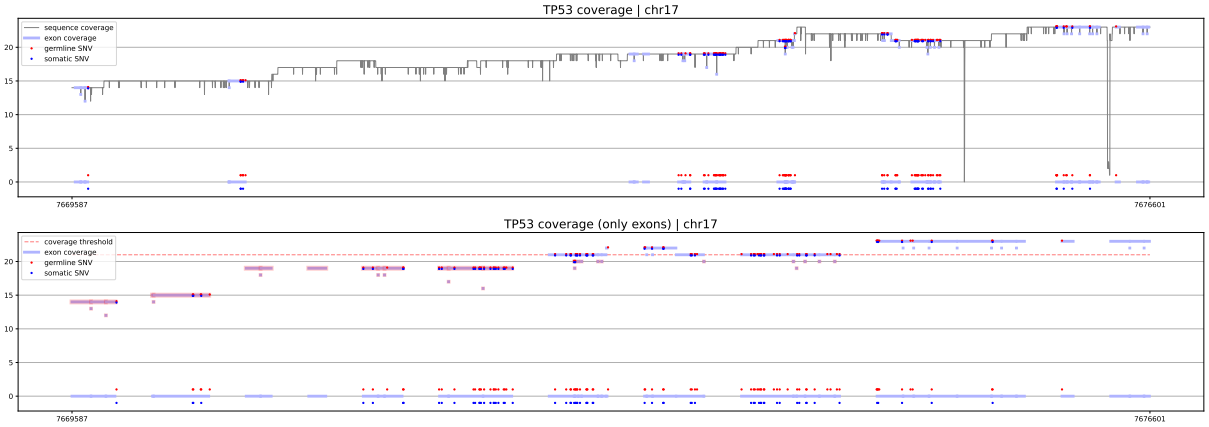
As the result, gene coverage report file and eight figures related to input genes are generated.

```
geneCoverage – HG003.pacbio–hifi.21x.haplotag.grch38.bam.per–base.bed

Input genes: ['MSH6', 'TP53', 'ABCA4', 'PDGFRB', 'LEMD2', 'CFTR', 'HTT', 'DMD']

Pathogenic (G – germline, S – somatic) SNV coverage:
– 'count' is the number of variants in a given region
– 'AT' is the percentage of SNV coverage above threshold (21x)
```

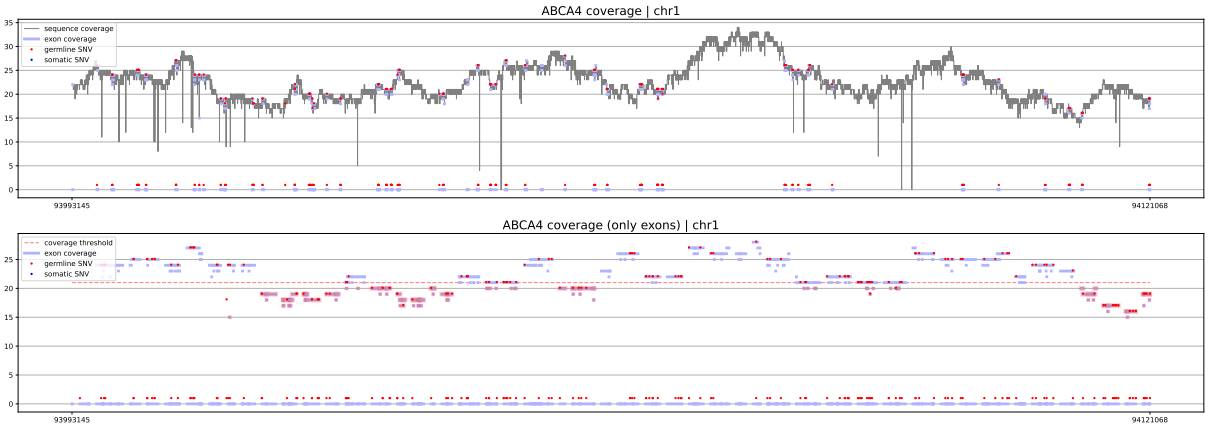
gene	chr	count(G)	median(G)	std(G)	min(G)	max(G)	AT(G)	count(S)	median(S)	std(S)	min(S)	max(S)	AT(S)
MSH6	chr2	118	30	4	14	33	93%	4	30	1	30	33	100%
TP53	chr17	89	21	2	14	23	65%	68	21	2	14	23	60%
ABCA4	chr1	166	22	3	16	28	68%	0	0	0	0	0	0%
PDGFRB	chr5	1	17	0	17	17	0%	1	17	0	17	17	0%
LEMD2	chr6	0	0	0	0	0	0%	0	0	0	0	0	0%
CFTR	chr7	192	27	4	19	39	94%	0	0	0	0	0	0%
HTT	chr4	0	0	0	0	0	0%	0	0	0	0	0	0%
DMD	chrX	144	11	4	5	18	0%	0	0	0	0	0	0%



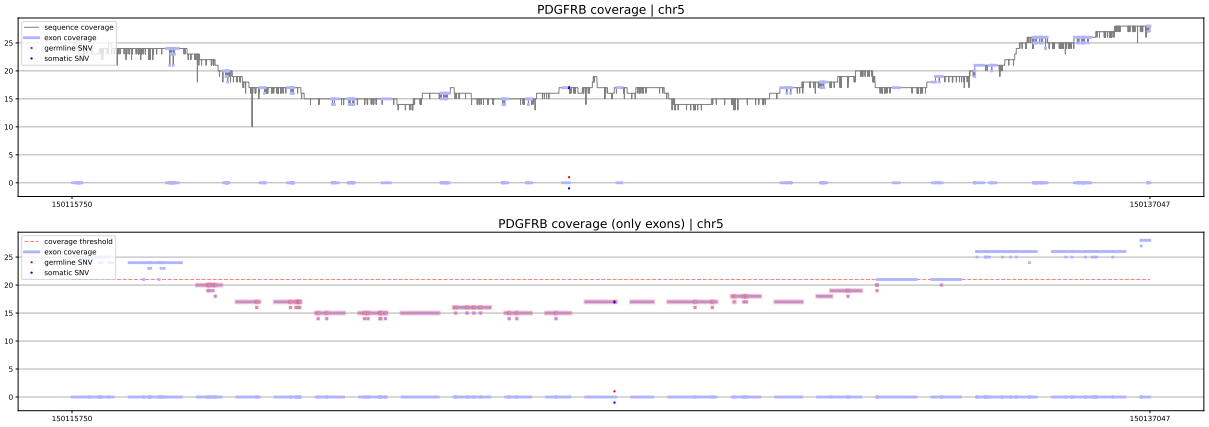
Supplementary Figure 2: MSH6 coverage



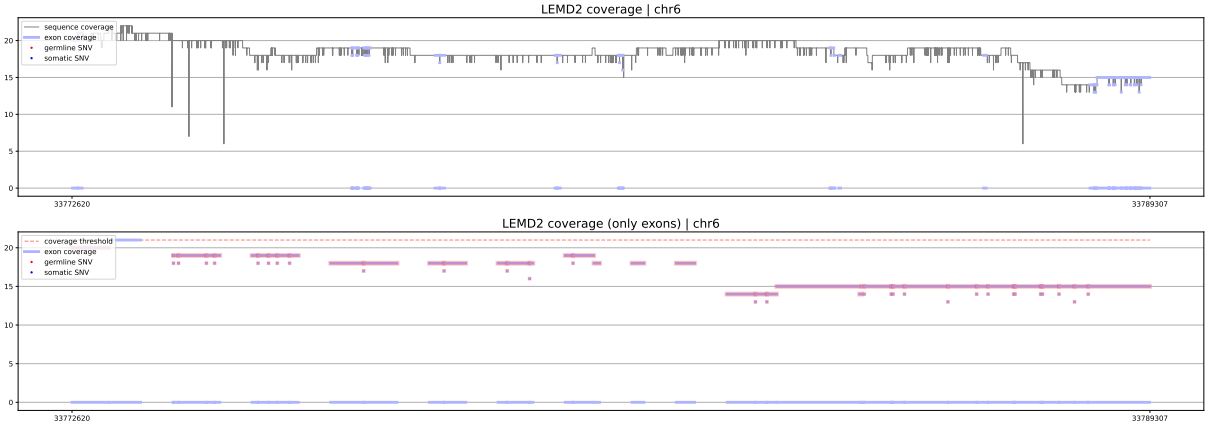
Supplementary Figure 3: TP53 coverage



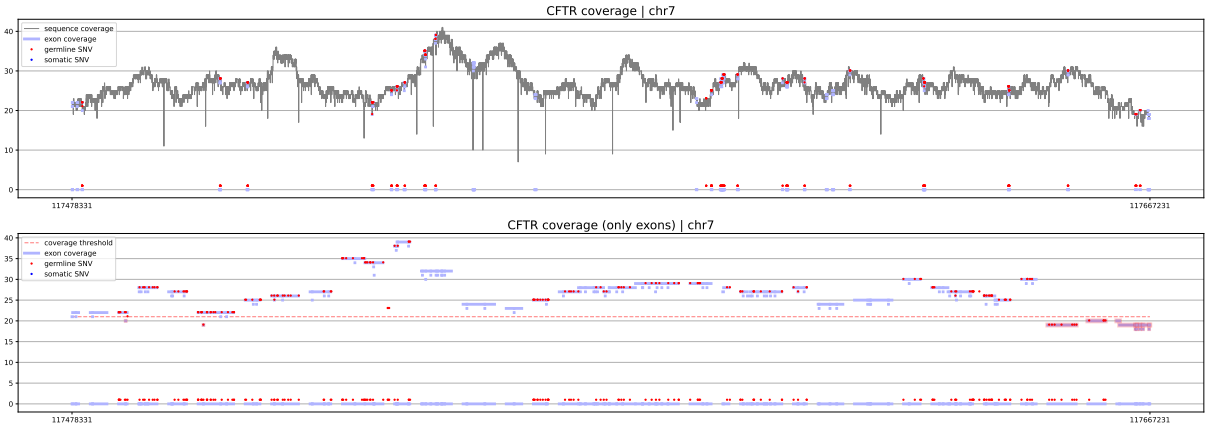
Supplementary Figure 4: ABCA4 coverage



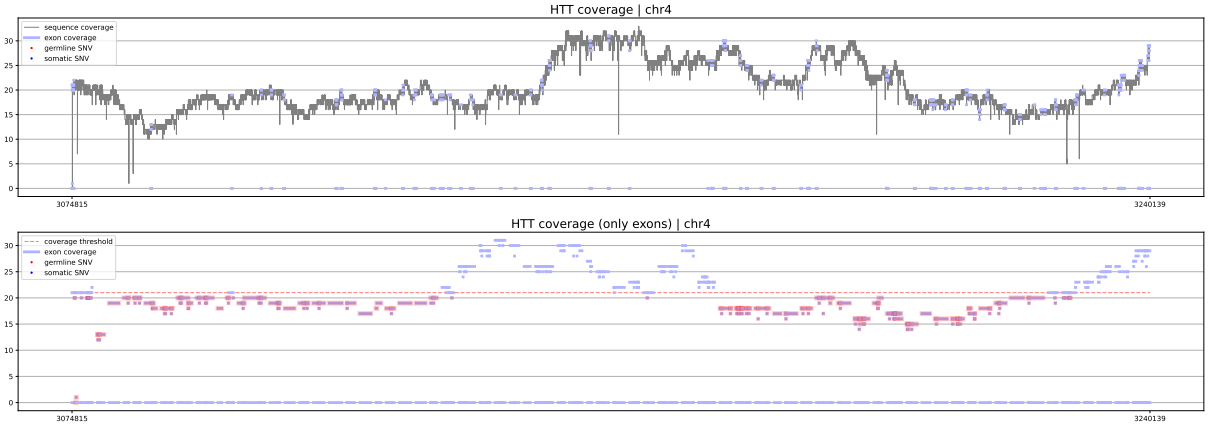
Supplementary Figure 5: PDGFRB coverage



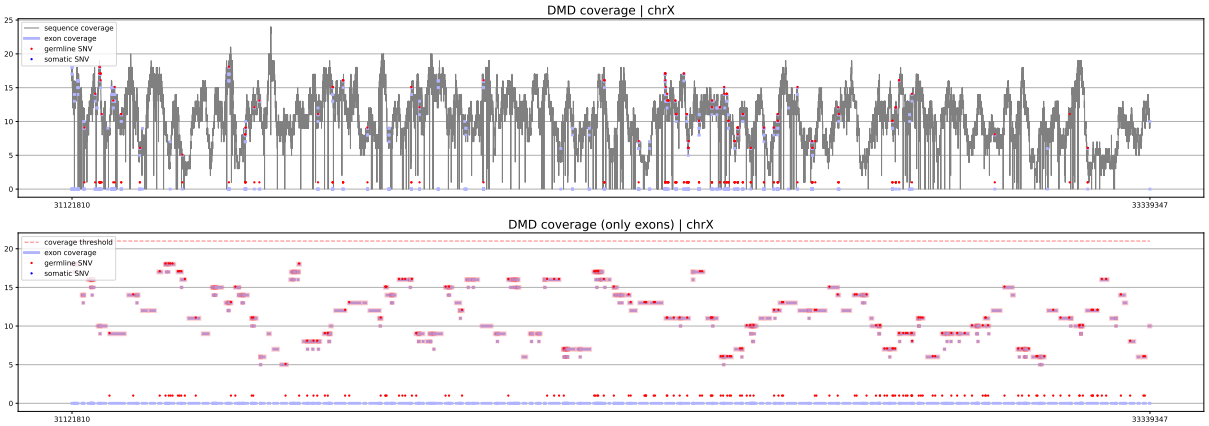
Supplementary Figure 6: LEMD2 coverage



Supplementary Figure 7: CFTR coverage



Supplementary Figure 8: HTT coverage



Supplementary Figure 9: DMD coverage

5 Example of Clinical Depth Coverage for the whole sequence samples

- PacBio Long Read

```
docker run -it --rm -v ~/data/pacbio_hifi/:/input -v ~/data/pacbio_hifi/:/output \
-v ~/data/refs/:/refs porebskis/exnverify:0.89b ./SNVScore.py \
input/HG003.pacbio-hifi.21x.haplotag.grch38.bam.per-base.bed \
refs/SNV_patho_germline.txt refs/SNV_patho_somatic.txt 15
```

```
SNV coverage report - HG003.pacbio-hifi.21x.haplotag.grch38.bam.per-base.bed

81% of all pathogenic germline SNVs and 91% of all pathogenic somatic SNVs are covered above threshold (15)

Whole genome coverage:
median mean  std 1st quartile 3rd quartile  min     max
    22   32  190           18           26    0   24444

Pathogenic (G - germline, S - somatic) SNV coverage:
('count' is the number of variants in a given region)
```

region	count(G)	median(G)	std(G)	min(G)	max(G)	count(S)	median(S)	std(S)	min(S)	max(S)
ALL	14157	20	6	1	43	310	20	5	9	36
chr1	1182	21	5	8	37	10	24	3	17	29
chr2	1118	22	6	5	39	33	20	5	12	33
chr3	774	20	5	8	37	40	21	6	10	33
chr4	270	20	7	7	41	6	16	2	15	19
chr5	538	21	5	5	36	10	18	4	15	29
chr6	511	22	6	4	43	0	0	0	0	0
chr7	717	22	5	6	43	13	26	4	21	34
chr8	335	23	6	8	37	0	0	0	0	0
chr9	479	20	6	2	37	1	27	0	27	27
chr10	362	20	6	9	36	32	20	7	9	36
chr11	966	19	6	6	36	9	14	6	11	27
chr12	709	22	5	10	38	12	23	3	16	25
chr13	910	17	5	9	37	26	22	5	11	29
chr14	324	23	4	9	34	6	21	4	13	27
chr15	533	21	5	8	35	4	17	1	17	20
chr16	643	19	6	3	40	1	17	0	17	17
chr17	1561	20	4	8	37	88	21	2	14	24
chr18	154	20	6	11	35	4	25	2	25	30
chr19	720	18	4	7	27	10	12	4	9	20
chr20	146	18	5	8	35	0	0	0	0	0
chr21	174	14	6	2	28	1	22	0	22	22
chr22	181	16	5	1	29	2	18	6	12	24
chrX	850	10	4	3	24	2	14	4	9	18
chrY	0	0	0	0	0	0	0	0	0	0

- Illumina WGS

```
docker run -dit --rm -v ~/data/hiseqx_wgs_pcr_free/:/input \
-v ~/data/hiseqx_wgs_pcr_free/:/output -v ~/data/refs/:/refs \
porebskis/exnverify:0.89b ./SNVScore.py \
input/HG003.hiseqx.pcr-free.20x.dedup.grch38.bam.per-base.bed \
refs/SNV_patho_germline.txt refs/SNV_patho_somatic.txt 15
```

```
SNV coverage report - HG003.hiseqx.pcr-free.20x.dedup.grch38.bam.per-base.bed

94% of all pathogenic germline SNVs and 98% of all pathogenic somatic SNVs are covered above threshold (15)

Whole genome coverage:
median mean  std 1st quartile 3rd quartile  min     max
    24   28  158           21           28    0  144498

Pathogenic (G - germline, S - somatic) SNV coverage:
('count' is the number of variants in a given region)
```

region	count(G)	median(G)	std(G)	min(G)	max(G)	count(S)	median(S)	std(S)	min(S)	max(S)
ALL	14157	24	6	2	54	310	24	6	9	41
chr1	1182	25	6	10	44	10	26	4	18	33
chr2	1118	24	5	5	54	33	25	4	15	32
chr3	774	24	5	9	41	40	24	6	14	37
chr4	270	24	6	10	43	6	26	3	21	32
chr5	538	25	5	7	40	10	27	5	16	34
chr6	511	24	5	10	44	0	0	0	0	0

chr7	717	24	6	10	45	13	25	1	23	29
chr8	335	24	5	12	39	0	0	0	0	0
chr9	479	24	5	13	46	1	19	0	19	19
chr10	362	25	6	10	39	32	23	7	10	39
chr11	966	24	6	8	44	9	23	4	20	33
chr12	709	23	5	11	43	12	28	5	14	30
chr13	910	25	5	12	43	26	24	4	16	34
chr14	324	24	5	12	43	6	20	6	17	33
chr15	533	25	6	10	42	4	26	8	16	33
chr16	643	24	5	11	42	1	28	0	28	28
chr17	1561	25	5	7	42	88	24	6	16	41
chr18	154	24	5	11	41	4	24	4	17	28
chr19	720	25	5	10	41	10	27	6	12	36
chr20	146	24	5	11	36	0	0	0	0	0
chr21	174	22	7	2	33	1	27	0	27	27
chr22	181	25	5	12	41	2	22	2	19	24
chrX	850	12	4	2	26	2	12	4	9	16
chrY	0	0	0	0	0	0	0	0	0	0

- Illumina Exome Agilent

```
docker run –dit ––rm –v ~/data/novaseq_wes_agilent/:/input \
–v ~/data/novaseq_wes_agilent/:/output –v ~/data/refs/:/refs \
porebskis/exnverify:0.89b ./SNVScore.py \
input/HG003.novaseq.wes–agilent.100x.dedup.grch38.per–base.bed \
refs/SNV_patho_germline.txt refs/SNV_patho_somatic.txt 100
```

```
SNV coverage report – HG003.novaseq.wes–agilent.100x.dedup.grch38.per–base.bed

73% of all pathogenic germline SNVs and 84% of all pathogenic somatic SNVs are covered above threshold (100)

Whole genome coverage:
median mean  std 1st quartile 3rd quartile  min     max
    18   53  106            8           71    0   51189

Pathogenic (G – germline, S – somatic) SNV coverage:
('count' is the number of variants in a given region)

region count(G)  median(G) std(G) min(G) max(G) count(S)  median(S) std(S) min(S) max(S)
------------------------------------------------------------------------------------------
   ALL    14157        141    100      0   1087      310        162    108     46    863
------------------------------------------------------------------------------------------
  chr1     1182        138    109      0   1087       10        146     39     82    179
  chr2     1118        143     98     19    956       33        181     95     46    432
  chr3      774        152     98      0    646       40        165     85     68    468
  chr4      270        122     73      0    551        6         88     19     75    133
  chr5      538        146    113      0    875       10        187     68    130    373
  chr6      511        145    154      0   1070        0          0      0      0      0
  chr7      717        139    114      0    798       13        132     61    124    280
  chr8      335        138     97     23    753        0          0      0      0      0
  chr9      479        125     83      0    519        1         78      0     78     78
 chr10      362        154    112     12    863       32        136    214     50    863
 chr11      966        122     84      9    869        9         95     39     66    201
 chr12      709        142     91     11    580       12        150     31    119    214
 chr13      910        198     89      2    772       26        173    159     48    772
 chr14      324        152     82      6    452        6        198     77     92    309
 chr15      533        139     82     30    740        4        185     54     90    221
 chr16      643        136     80      7    602        1        226      0    226    226
 chr17     1561        163     91      0    938       88        197     71     88    359
 chr18      154        142     78      3    576        4        188     51    168    297
 chr19      720        148    105     11    644       10        120     72     54    280
 chr20      146        142    177      0   1084        0          0      0      0      0
 chr21      174        104     63     19    455        1         69      0     69     69
 chr22      181        132     55     33    301        2        237      9    228    246
  chrX      850         72     55      0    316        2        167     23    144    190
  chrY        0          0      0      0      0        0          0      0      0      0
```