Comparison of two target sequencing approaches. *

Dylan Lawless, PHD[1]

[1]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland, Dylan.Lawless@epfl.ch

1 Introduction

The two datasets (AH S1 and CH S2) of paired-end short reads were generated from the same human DNA NGS library for clinical diagnosis of solid tumor. They were obtained through two different target sequencing approaches with corresponding target region file provided (hg19). Herein, we evaluate their performances for use in clinical diagnose.

2 Details for technical writing report

Note: This section contains the main information required for a technical report. The remainder of this document contains additional information which is *not* required for further reports, and includes duplicate references to to files listed here. Please include:

- **Introduction**: all details from sec. 1.
- **Data source**: all details from sec. 3.1.
- **Protocol details**: a hyperlink to the protocol source: [link].
- **Result**: Result summary sec. 4.1.
- **Figures**: 1, 7, and 10.
- **Conclusion**: all details from sec 5.

3 Methods

3.1 Data source

- **Description**: The sequencing group performed library preparation for two (2) samples (AH and CH) using Kit X. The performance of each was assessed with several bioinformatic methods including read quality control and performance when aligning to reference genome [code available here].
- **Data source**: raw sequence data was received from [link group and contact address].
- **Date**: 2022-02-14
- **Link to tickets where submission was logged**: [link]
- **Data integrity**: [link metadata md5sum].

*This document's source code is available from the GitHub repository. All code used in this report is available on the GitHub repository.

3.2 Configuration

Local env: macOS v11.6 software, Remote env: Red Hat Enterprise Linux Server 7.6 (Maipo) software, compiler intel: (19.0.5 and 19.1.1), software, R: R v4.1.0 Camp Pontanezen software, R libraries: versions unlisted, fastqc: v0.11.7 software, samtools: v1.10 software, bwa: v0.7.17 software, picard: v2.20.8 software. qualimap v2.2.1 qualimap.

3.3 Analysis code

All code available at link: GitHub/kit_assess.
Protocol order:

- src/md5sum.ch
- src/read_count.sh
- fastQC: manual run all fastq and save to ./processed/fastqc
- fastqcr: Assess fastqc rerpots further
- src/target_info.sh
- src/1.trim.sh
- src/2.align.sh
- src/3.sort.sh
- src/mapping.sh

4 Results

Data for this report can be found on link: GitHub/kit_assess. Analysis results are separated into QC of fastq data (sec. 4.2), QC after alignment to reference genome GRCh37 (sec. 4.3), and exploration of target coding regions (sec. 4.4).

4.1 Result summary

Based on [1] QC of fastq data, [2] QC after alignment to reference genome GRCh37, and [3] exploration of target coding regions, AH S1 performed better than CH S2 for use in clinical diagnosis of solid tumor. The read depth and uniform coverage provided for AH S1 outperformed CH S2, and is beneficial in tumor sequencing for calling germline variants, somatic variants, CNV or structural changes.

4.2 Fastq data

To assess the qulity of fastq data, FastQC was used. Full html reports for each file are linked:

- AH_S1_L001_R1_fastqc
- AH_S1_L001_R2_fastqc
- CH_S2_L001_R1_fastqc
- CH_S2_L001_R2_fastqc

The results of fastQC were also assessed by use of fastqcr. The full html report is linked:

- Report assessment of fastQC

1. Total sequences or the number of reads for all samples: 1000000
2. Per base sequence quality: All samples perform sufficiently. Summarised in Figure 1
    - Median value (red line): **AH good quality** (qual >28), **CH good quality** (qual >28).
    - Inter-quartile range (25-75%) (yellow box): **AH good quality** (qual >28), **CH medium to good quality** (qual >20).
    - Upper and lower 10% and 90% whiskers points: **AH medium to good quality** (qual >28), **CH poor to good quality** (qual >14).
    - Mean quality (blue line): **AH good quality** (qual >28), **CH medium to good quality** (qual >20).
3. Per tile sequence quality: **All samples perform sufficiently**. No warning.
4. Per sequence quality score: **All samples perform sufficiently**, summarised in Figure 2.
5. Per base sequence content. **All samples flagged** with warning indicating a difference greater than 10% in any position. However, this is potentially due targeted capturing, Figure 3.
6. Per sequence GC content. **All samples fail** based on modal GC content as calculated from the observed data and used to build a reference distribution. The sum of the deviations from the normal distribution represents more than 30% of the reads. However, the sharp peaks (as seen on AH [Top]) are most likely due to enriched duplicate sequences from targeted capturing and do not necessarily indicate poor quality, Figure 4.
7. Sequence Length Distribution: **AH reads were all 150**, while **CH reads were 35-151**.
8. Sequence Duplication Levels: Percentage of duplicate reads were **AH 96.01%-96.55%** and **CH 65.44%-67.13%**. Figure 5
9. Adapter Content: Figure 6, **AH Illumina Universal adaptor** and **CH Adaptor not detected**.

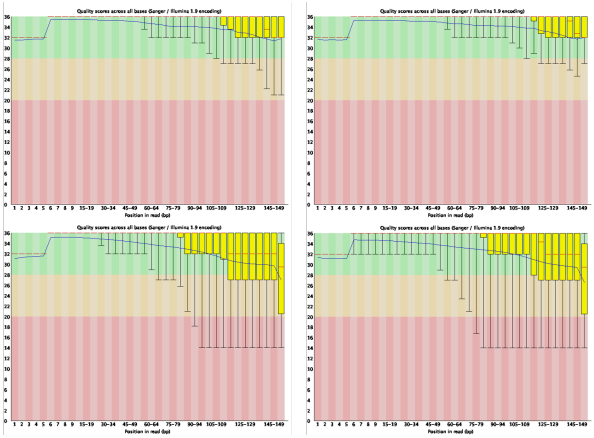

Figure 1: Per base sequence quality score: [Top] AH [Bottom] CH. AH outperforms CH for both reads. Central red line shows the median value. Inter-quartile range 25-75% (yellow box) Upper and lower whiskers represent the 10% and 90% points. Mean quality (blue line).

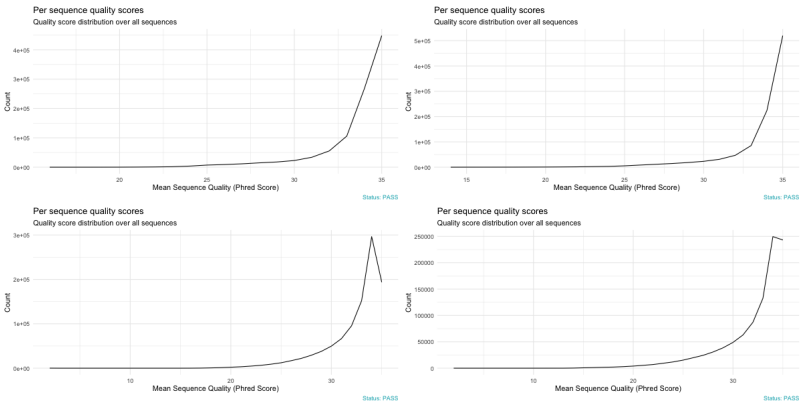

Figure 2: Per sequence quality score: [Top] AH [Bottom] CH. All samples perform sufficiently. AH outperforms CH for both reads. Most frequently observed mean quality is above 30 for all samples; less than 0.2% error rate.

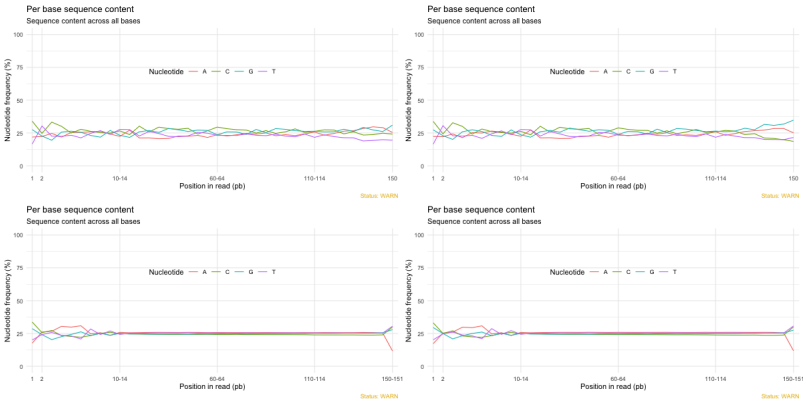

Figure 3: Per base sequence content: [Top] AH [Bottom] CH.

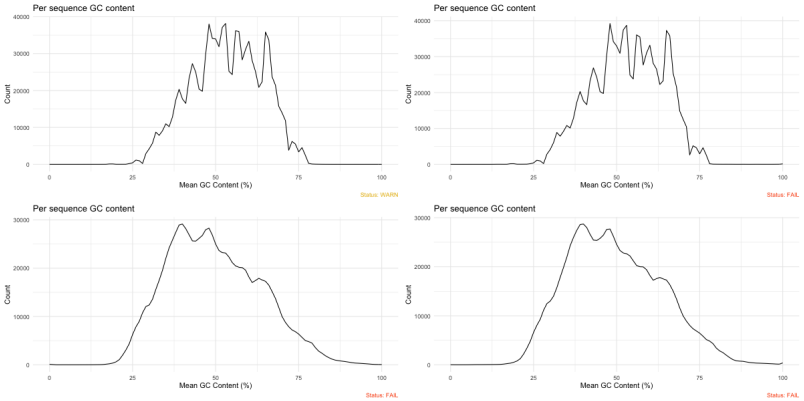

Figure 4: Per sequence GC content: [Top] AH [Bottom] CH. A normal distribution is expected for genomic DNA. However, targeted sequencing may produce difference distributions.

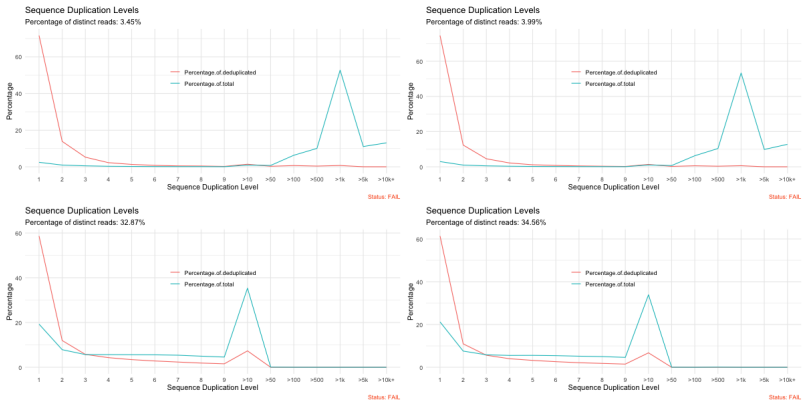

Figure 5: Sequence Duplication Levels: [Top] AH [Bottom] CH. The plot shows the proportion of the library which is made up of sequences in each of the different duplication level bins. There are two lines on the plot. The blue line takes the full sequence set and shows how its duplication levels are distributed. In the red plot the sequences are de-duplicated and the proportions shown are the proportions of the deduplicated set which come from different duplication levels in the original data. In a properly diverse library most sequences should fall into the far left of the plot in both the red and blue lines. The fail status is due to non-unique sequences making up more than 50% of the total, which is not a problem for targeted capture libraries.

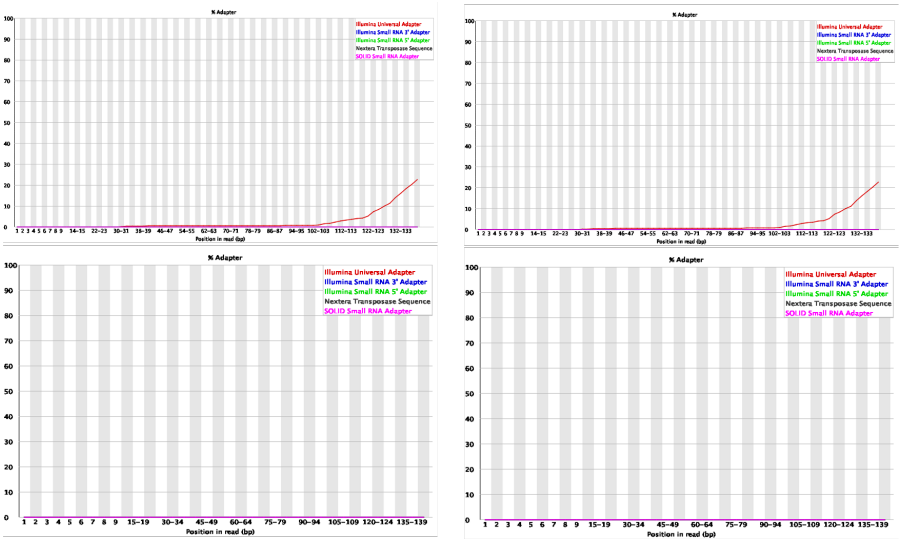

Figure 6: Adapter content plot: [Top] AH [Bottom] CH. Acumulative percentage count of the proportion of library which has seen each of the adapter sequences at each position.

Table 1: Samtools flagstat mapping summary. Alignment with GRCh37, sorted reads.

CH	AH	
2011262	1999498	in total (QC-passed reads & + QC-failed reads)
15710	612	secondary
0	0	supplementary
0	0	duplicates
2006501	1997929	mapped (99.76% : N/A, 99.92% : N/A)
1995552	1998886	paired in sequencing
997776	999443	read1
997776	999443	read2
1968314	1992738	properly paired (98.64% : N/A, 99.69% : N/A)
1986886	1996840	with itself and mate mapped
3905	477	singletons (0.20% : N/A, 0.02% : N/A)
14488	1612	with mate mapped to a different chr
8748	1426	with mate mapped to a different chr (mapQ>=5)

4.3 Alignment data

Fastq were trimmed using TrimGalore. Trim Galore also requires the use of cutadapt.

Reads were aligned to GRCh37 using BWA MEM and converted to bam format with samtools.

The aligment data was assessed using

- samtools flagstat: get mapping summary.
- samtools depth: read depth for at all positions of the reference genome, e.g. how many reads are overlapping the genomic position.
- qualimap: examines sequencing alignment data in SAM/BAM files according to the features of the mapped reads and provides an overall view of the data that helps to the detect biases in the sequencing and/or mapping of the data and eases decision-making for further analysis.

Qualimap full report link: sample AH

Qualimap full report link: sample CH

1. Samtools flagstat mapping summary shows alignment with GRCh37, sorted reads, Table 1.
2. Mapping quality histogram indicate AH performing better than CH, Figure 7.
3. Genome coverage histogram shows that AH produced a normal distribution of coverage depths while CH had an enrichment for some genomic regions, Figure 8.
4. The duplication rate histograms are shown in Figure 9.
5. Genome coverage across GRCh37 shows a uniform distribution of reads for AH [Top], while CH [Bottom] has high depth in some regions with lower coverage in others, Figure 10.

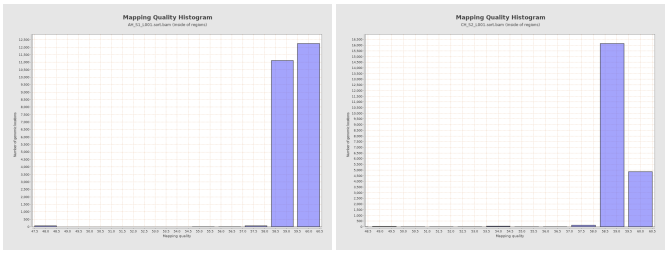

Figure 7: Mapping quality histogram (GRCh37). [Left] AH [Right] CH. Histogram of the number of genomic locations having a given mapping quality. To construct the histogram mean mapping quality is computed at each genome position with non-zero coverage and collected. According to Specification of the SAM format the range for the mapping quality is [0-255].

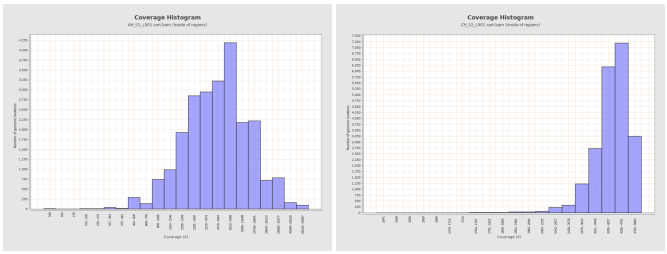

Figure 8: Genome coverage histogram (GRCh37). [Left] AH [Right] CH. Histogram of the number of genomic locations having a given coverage rate. The bins of the x-axis are conveniently scaled by aggregating some coverage values in order to produce a representative histogram also in presence of the usual NGS peaks of coverage.

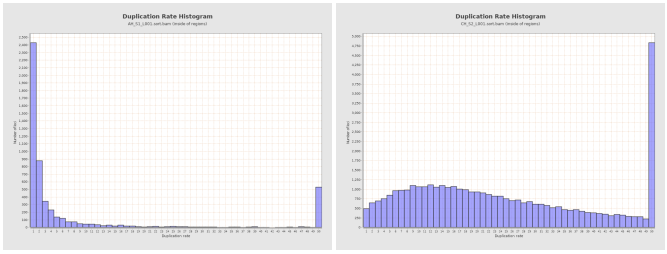

Figure 9: Duplication rate histogram (GRCh37). [Left] AH [Right] CH. This plot shows the distribution of duplicated read starts. Due to several factors (e.g. amount of starting material, sample preparation, etc) it is possible that the same fragments are sequenced several times. For some experiments where enrichment is used (e.g. ChIP-seq) this is expected at some low rate. If most of the reads share the exact same genomic positions there is very likely an associated bias.

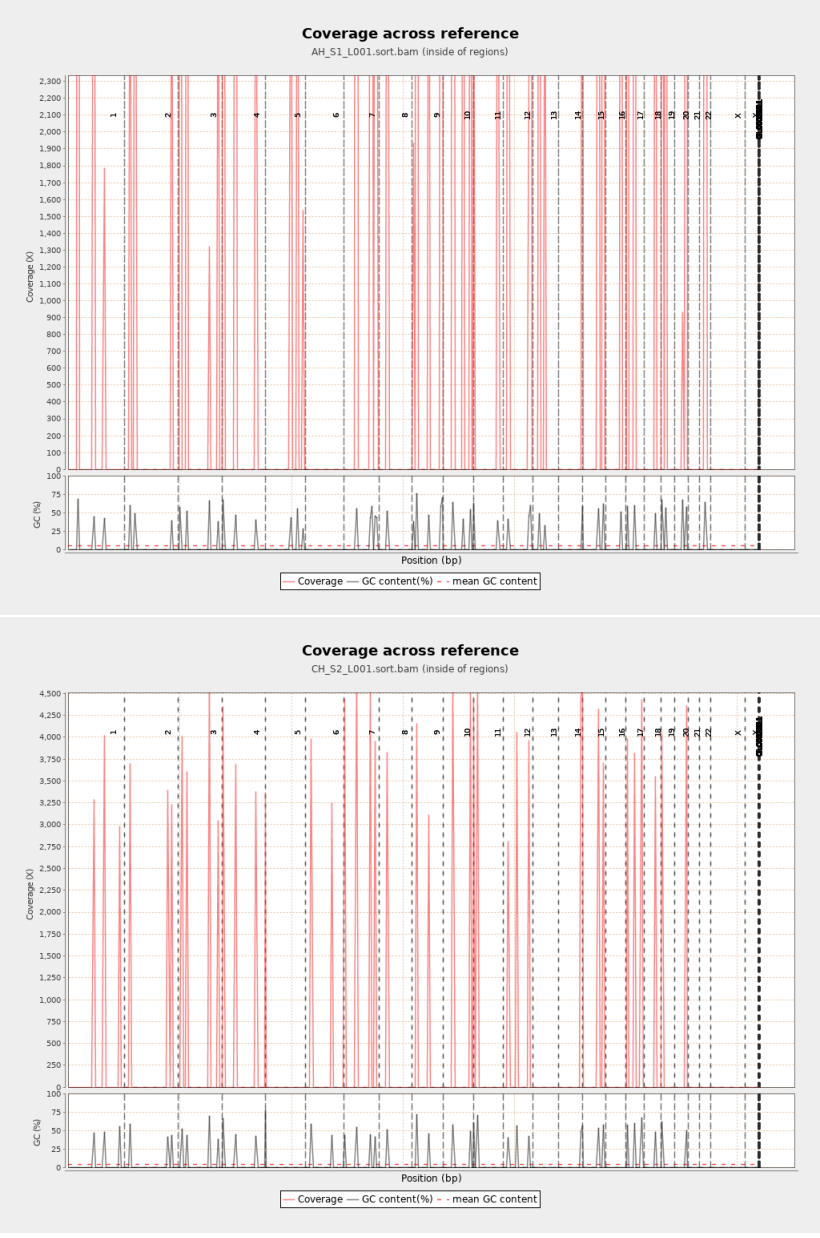

Figure 10: Genome coverage across reference (GRCh37). [Top] AH [Bottom] CH. Each plot consists of two figures. The upper figure provides the coverage distribution (red line) and coverage deviation across the reference sequence. The coverage is measured in X number of reads. The lower figure shows GC content across reference (black line) together with its average value (red dotted line).

4.4 Target coding genes

To learn about what target regions are captured in this dataset, the coordinates (hg19) were annotated using Ensembl Biomart. The data was annotated using grch37.ensembl.org/biomart dataset Ensembl Genes 105 (GRCh37.p13). Output Attributes queried for these regions: Gene stable ID, Transcript stable ID, Gene name, Gene start (bp), Gene end (bp), Chromosome/scaffold name, Gene % GC content, HGNC symbol, HGNC ID. The gene symbols for each target list are shown in Table 2:

- Target coding genes table (html version).

Table 2: Targeted coding genes in GRCh37 present in for each methods.

Chr	Start	End	Gene name	pc GC content	AH S1	CH S2
1	43803478	43818443	MPL	48.47	Yes	No
1	115247090	115259515	NRAS	39	Yes	Yes
1	162601163	162757190	DDR2	40.47	Yes	Yes
2	25455845	25565459	DNMT3A	53.5	Yes	No
2	29415640	30144432	ALK	43.51	Yes	Yes
2	47922669	48037240	MSH6	44.65	Yes	No
2	48016455	48132932	FBXO11	39.46	Yes	No
2	209100951	209130798	IDH1	41.27	Yes	Yes
2	212240446	213403565	ERBB4	34.69	Yes	Yes
3	10182692	10193904	VHL	47.55	Yes	No
3	37034823	37107380	MLH1	42.01	Yes	No
3	41236328	41301587	CTNNB1	39.91	Yes	Yes
3	138663066	138665982	FOXL2	64.62	Yes	Yes
3	178865902	178957881	PIK3CA	35.78	Yes	Yes
4	1795034	1810599	FGFR3	65.82	Yes	Yes
4	54243810	55161439	FIP1L1	40.97	Yes	Yes
4	55095264	55164414	PDGFRA	43.65	Yes	Yes
4	55524085	55606881	KIT	40.87	Yes	Yes
4	55919227	55958701	RP11-530I17.1	40.25	Yes	No
4	55944644	55991756	KDR	40.69	Yes	No
4	153242410	153457253	FBXW7	35.41	Yes	Yes
4	153258807	153259250	RP11-461L13.2	34.91	Yes	No
5	112043195	112181936	APC	37.63	Yes	No
5	112162910	112203279	CTC-554D6.1	39.75	Yes	No
5	149432854	149492935	CSF1R	48.1	Yes	No
5	170814120	170838141	NPM1	43.6	Yes	No
7	55086714	55324313	EGFR	45.08	Yes	Yes
7	55247443	55256627	EGFR-AS1	50.41	Yes	Yes
7	116312444	116438440	MET	39.03	Yes	Yes

7	128828713	128853386	SMO	49.17	Yes	No
7	128849616	128853386	RP11-286H14.8	57.23	Yes	No
7	140419127	140624564	BRAF	37.97	Yes	Yes
7	148504475	148581413	EZH2	39.21	Yes	No
8	38268656	38326352	FGFR1	50.55	Yes	Yes
8	38279407	38283614	RP11-350N15.4	50.78	Yes	No
9	4985033	5128183	JAK2	37.53	Yes	No
9	5077163	5084580	AL161450.1	34.29	Yes	No
9	21802635	22032985	RP11-145E5.5	40.29	Yes	Yes
9	21967751	21995300	CDKN2A	41.19	Yes	Yes
9	80331003	80646374	GNAQ	39.13	Yes	Yes
9	133589333	133763062	ABL1	44.47	Yes	No
9	135766735	135820020	TSC1	42.78	Yes	No
9	139388896	139440314	NOTCH1	63.39	Yes	No
10	43572475	43625799	RET	55.73	Yes	Yes
10	89622870	89731687	PTEN	35.77	Yes	No
10	123237848	123357972	FGFR2	45.41	Yes	Yes
11	532242	537287	HRAS	69.01	Yes	Yes
11	108093211	108239829	ATM	37.52	Yes	No
11	108179246	108338258	C11orf65	38.99	Yes	No
12	25357723	25403870	KRAS	36.37	Yes	Yes
12	112856155	112947717	PTPN11	43.16	Yes	Yes
12	121416346	121440315	HNF1A	52.2	Yes	No
13	28577411	28674729	FLT3	43.22	Yes	No
13	48877887	49056122	RB1	36.91	Yes	No
14	105235686	105262088	AKT1	64.08	Yes	Yes
15	66679155	66784650	MAP2K1	44.72	Yes	Yes
15	66782086	66784447	CTD-3185P2.1	44.37	Yes	No
15	66782473	66790151	SNAPC5	44.54	Yes	No
15	90626277	90645736	IDH2	52.54	Yes	Yes
16	68771128	68869451	CDH1	46.89	Yes	No
17	7565097	7590856	TP53	48.85	Yes	Yes
17	37844167	37886679	ERBB2	52.09	Yes	Yes
18	48494389	48584514	RP11-729L2.2	41.11	Yes	No
18	48494410	48611415	SMAD4	40.04	Yes	Yes
19	1189406	1228428	STK11	58.47	Yes	No
19	3094408	3124002	GNA11	61.66	Yes	Yes
19	3118663	3119302	AC005262.3	64.69	Yes	Yes
19	17935589	17958880	JAK3	54.93	Yes	No
20	35973088	36034453	SRC	54.1	Yes	No

20	57414773	57486247	GNAS	47.09	Yes	Yes
22	24129150	24176703	SMARCB1	50.77	Yes	No

5 Conclusion

[1] Based on QC of fastq data (sec. 4.2), **AH S1 outperformed CH S2**. [2] Based on QC after alignment to reference genome GRCh37 (sec. 4.3), **AH S1 outperformed CH S2.** However, it should be noted that non-uniform coverage favouring a subset of genes might be preferable in some circumstances; additional information on study design required. [3] Based on coverage for coding sequences within targeted regions (sec. 4.4), **AH S1 outperformed CH S2**. However, additional information on study design required for confirmation

6 Colophon

This document style is derived from the web design for SOPHiA genetics, using fonts Source Sans Pro (light, regular, and italic). Font colors are set as “sophiablue” #0A2E4A and “sophiapink” #E32A5C. Writing was done using LaTeX and html pages are hosted on https://lawlessgenomics.com. For contact: see here.