KINGHORN CENTRE FOR CLINICAL GENOMICS
GARVAN INSTITUTE OF MEDICAL RESEARCH

Garvan Institute of Medical Research
384 Victoria Street
Darlinghurst, NSW 2010
Australia
T: + 61 (0)2 9355 5846
E: kccgseq@garvan.org.au

KCCG Sequencing Project Report v1.5

Project Information

Item	Value
Project Name	R_150619_JOHDOE_FGS
Manifest	R_150619_JOHDOE_FGS_M001
External Project Name	My Sequencing Project
Name	Dr John Doe
Email	j.doe@email.org.au
Address	123 Long Street South Bay, NSW 2999 Australia
Report Date	2015-07-19
Number of Samples	3
Lanes Used	5

Each sample provided is assigned an internal ID for tracking and processing. This is a unique ID and is referred to throughout this document as “SampleID”. The following table maps these internal IDs to the sample ID provided to KCCG during the sample submission (ExternalID).

	SampleID	ExternalID
3	150622_FR07889450	Blood_DNA
2	150622_FR07889451	Tumour_DNA_1
1	150622_FR07889465	Tumour_DNA_2

Overall Summary

The table below summarises the overall performance of each sample. It provides the raw sequencing Yield (Megabases), the percentage of bases that had quality >= Q30, and compares the requested sequencing coverage to the mean sequencing coverage that we achieved . It is important to note that to determine coverage statistics, we aligned the sequencing reads to the reference genome using a fast genome aligner (iSAAC; see methods). Your values may differ if you use a different short-read aligner.

	SampleID	Yield (Mb)	% Bases >= Q30	Requested Coverage	Mean Coverage (Illumina)
1	150622_FR07889450	132280	84.44	Human WGS 30x (HiSeq X) v2.0	39.36
2	150622_FR07889451	283864	85.44	Human WGS 60x (HiSeq X) v2.0	83.89
3	150622_FR07889465	270355	84.13	Human WGS 60x (HiSeq X) v2.0	76.34

Samples Information

Samples are sequenced on multiple lanes in order to achieve the requested sequencing coverage and quality. In some cases, even a 30x genome requires more than one sequencing lane. We refer to each lane that a sample is sequenced on as an Indivisible Unit of Sequencing (IUS).

Each IUS is named using the following pattern: FlowCellID_Lane_SampleID_Species_Index_Manifest

Item	Description
FlowCellID	The ID of the flow cell used for sequencing
Lane	The lane on the flow cell used for the sample
SampleID	The sample ID consist of the date of submition and Sample name
Species	The species the sample was taken from (default is human)
Index	The index sequence for multiplexing (not currently used for HiSeq X)
Manifest	Manifest consist of the project name followed by the manifest id

The following table shows the samples for each IUS.

	IUS	FlowCellID	Lane	SampleID	Species	Index	Manifest
1	H75LYCCXX_3_150622_FR07889450_Human__R_150619_JOHDOE_FGS_M001	H75LYCCXX	3	150622_FR07889450	Human		R_150619_JOHDOE_FGS_M001
2	H72YGCCXX_3_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	H72YGCCXX	3	150622_FR07889451	Human		R_150619_JOHDOE_FGS_M001
3	HCCJFCCXX_8_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	HCCJFCCXX	8	150622_FR07889451	Human		R_150619_JOHDOE_FGS_M001
4	H72YGCCXX_1_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	H72YGCCXX	1	150622_FR07889465	Human		R_150619_JOHDOE_FGS_M001
5	HCCJFCCXX_7_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	HCCJFCCXX	7	150622_FR07889465	Human		R_150619_JOHDOE_FGS_M001

Library and Read Specifications

Below are statistics that describe the short-read sequencing library prepared from the sequenced samples. Sequencing read type and read length(s) are also indicated

Read Type: Paired

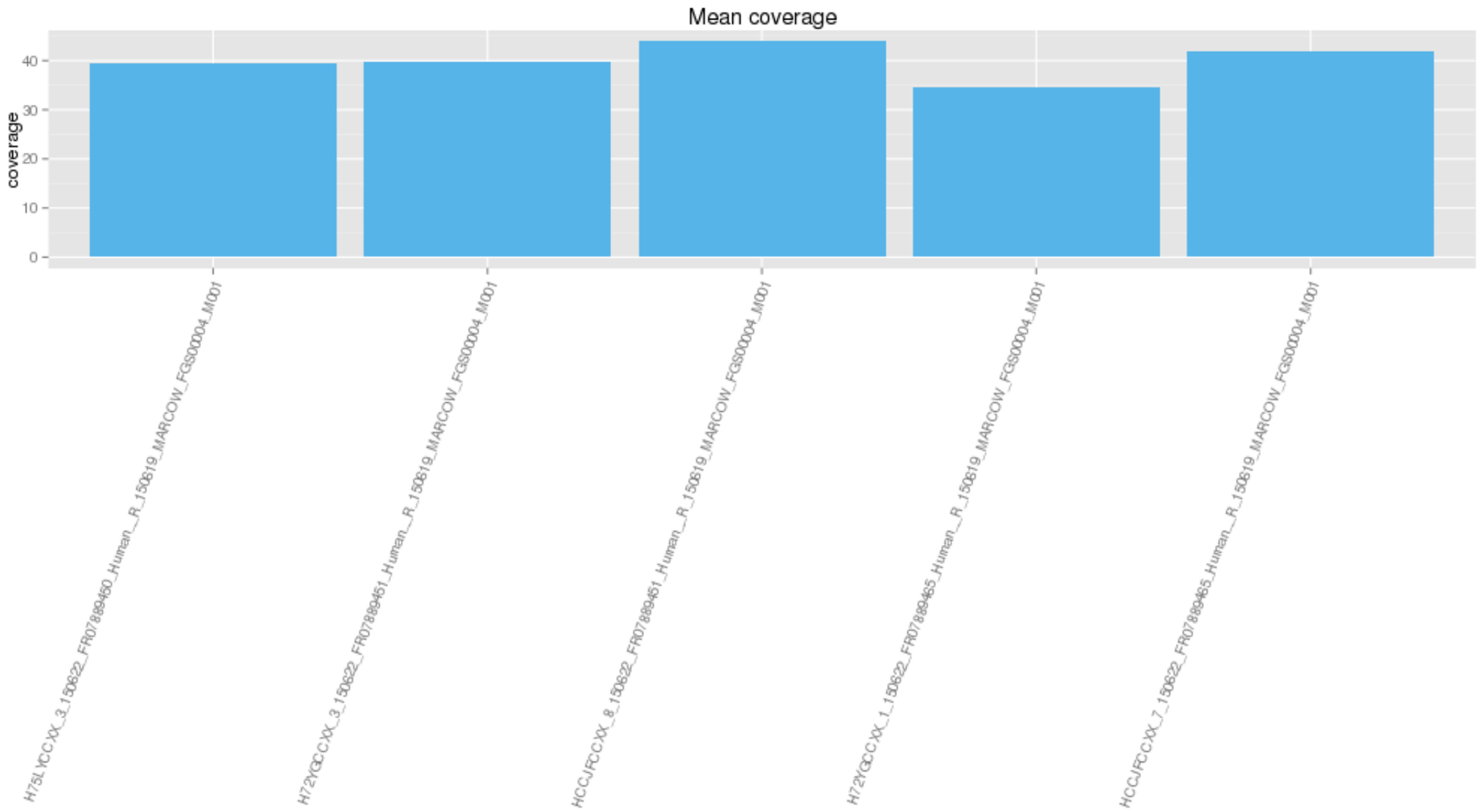
Read 1 Length: 150

Read 2 Length: 150

Data Quality and Coverage

The following table summarises the performace of each indivisible unit of sequencing (IUS).

	IUS	Genome Territory	Mean Coverage (Illumina)	Yield (MB)	% >= Q30
1	H75LYCCXX_3_150622_FR07889450_Human__R_150619_JOHDOE_FGS_M001	2900434419	39.36	132280	84.44
2	H72YGCCXX_3_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	2900434419	39.81	135699	85.61
3	HCCJFCCXX_8_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	2900434419	44.07	148165	85.28
4	H72YGCCXX_1_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	2900434419	34.51	127496	83.57
5	HCCJFCCXX_7_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	2900434419	41.83	142859	84.62

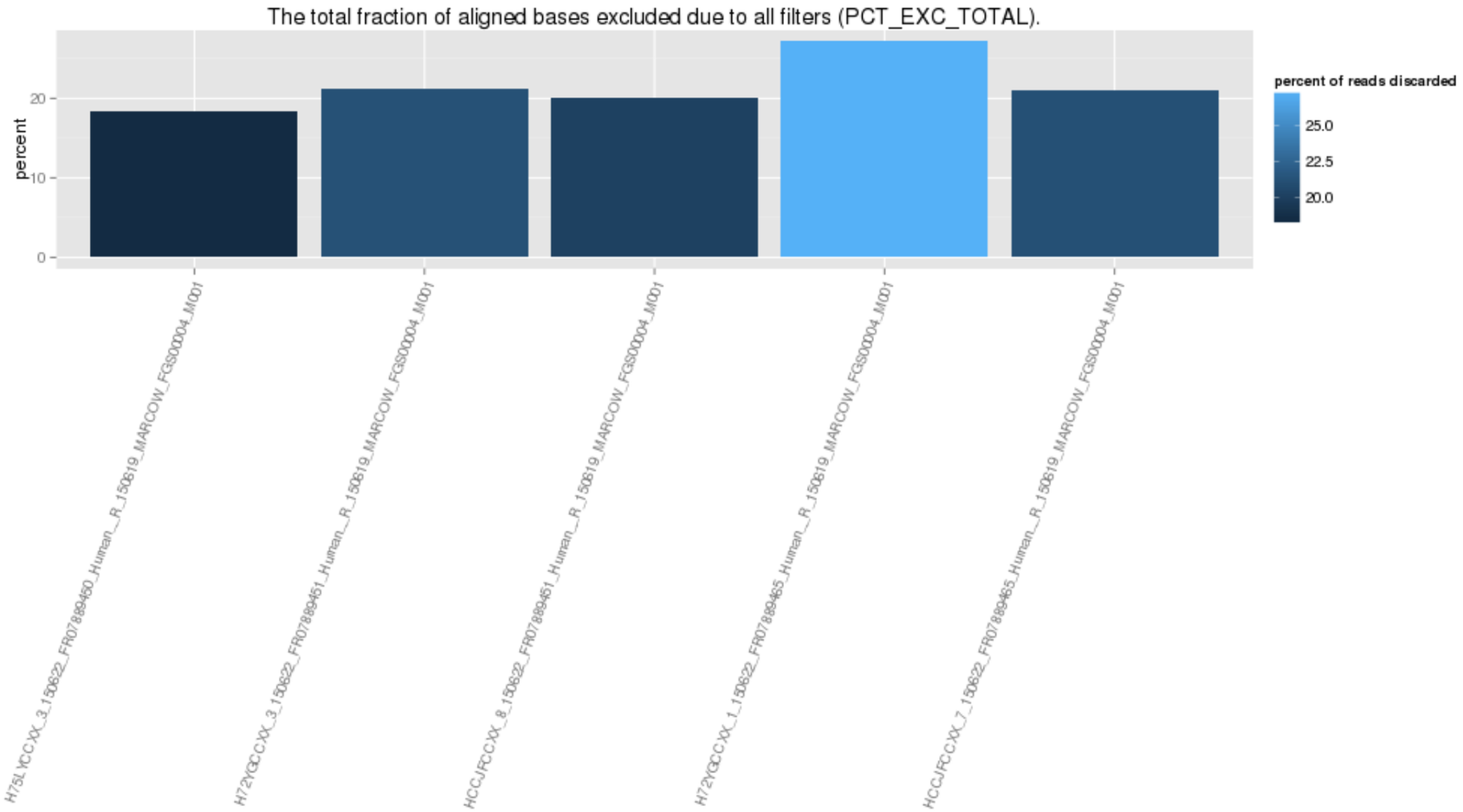


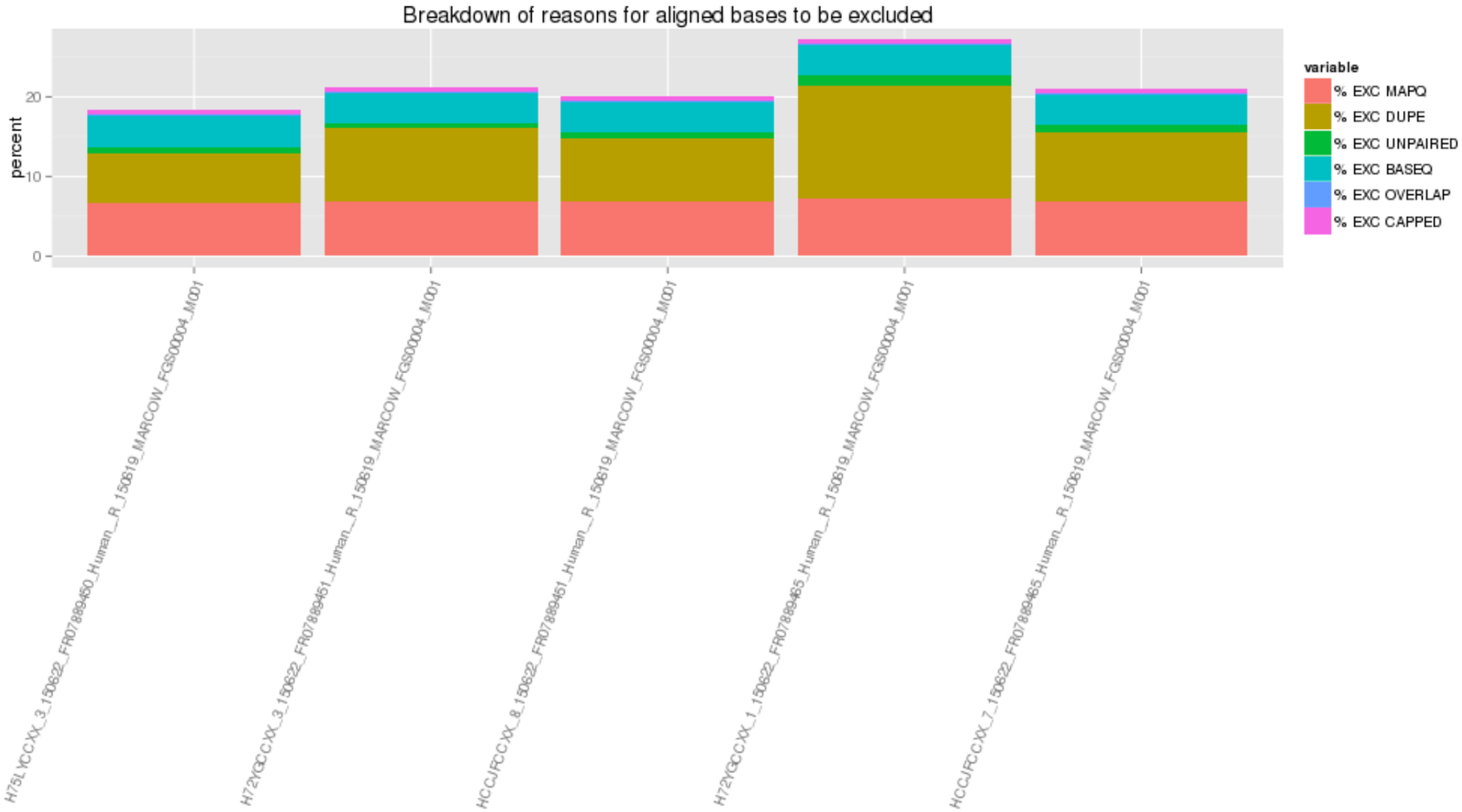
Poor Quality Reads

Some bases are filtered out from the alignments. This section shows a detailed breakdown of the criteria to exclude data and the percentage of reads excluded in each category.

> PCT_EXC_BASEQ: The fraction of aligned bases that were filtered out because they were of low base quality (default is < 20).
>
> PCT_EXC_CAPPED: The fraction of aligned bases that were filtered out because they would have raised coverage above the capped value (default cap = 250x).
>
> PCT_EXC_DUPE: The fraction of aligned bases that were filtered out because they were in reads marked as duplicates.
>
> PCT_EXC_MAPQ: The fraction of aligned bases that were filtered out because they were in reads with low mapping quality (default is < 20).
>
> PCT_EXC_OVERLAP: The fraction of aligned bases that were filtered out because they were the second observation from an insert with overlapping reads.

	IUS	% EXC TOTAL	% EXC MAPQ	% EXC DUPE	% EXC UNPAIRED	% EXC BASEQ	% EXC OVERLAP	% EXC CAPPED
1	H75LYCCXX_3_150622_FR07889450_Human__R_150619_JOHDOE_FGS_M001	18.39	6.682	6.238	0.7766	3.967	0.1104	0.6171
2	H72YGCCXX_3_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	21.15	6.910	9.269	0.5603	3.726	0.1067	0.5801
3	HCCJFCCXX_8_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	20.06	6.906	7.847	0.7094	3.880	0.1061	0.6114
4	H72YGCCXX_1_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	27.27	7.231	14.235	1.1631	3.890	0.1624	0.5887
5	HCCJFCCXX_7_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	21.08	6.918	8.699	0.8230	3.871	0.1116	0.6615


The total fraction of aligned bases excluded due to all filters (PCT_EXC_TOTAL).
percent
20
10
0
percent of reads discarded
25.0
22.5
20.0
50822_FR07889450_Human_R_150819_MARCOW_FGS00004_M001
50822_FR07889451_Human_R_150819_MARCOW_FGS00004_M001
50822_FR07889451_Human_R_150819_MARCOW_FGS00004_M001
50822_FR07889465_Human_R_150819_MARCOW_FGS00004_M001
50822_FR07889465_Human_R_150819_MARCOW_FGS00004_M001
H75LYCCXX_3_1
H72YGCCXX_3_1
HCCJFCCXX_8_1
H72YGCCXX_1_1
HCCJFCCXX_7_1

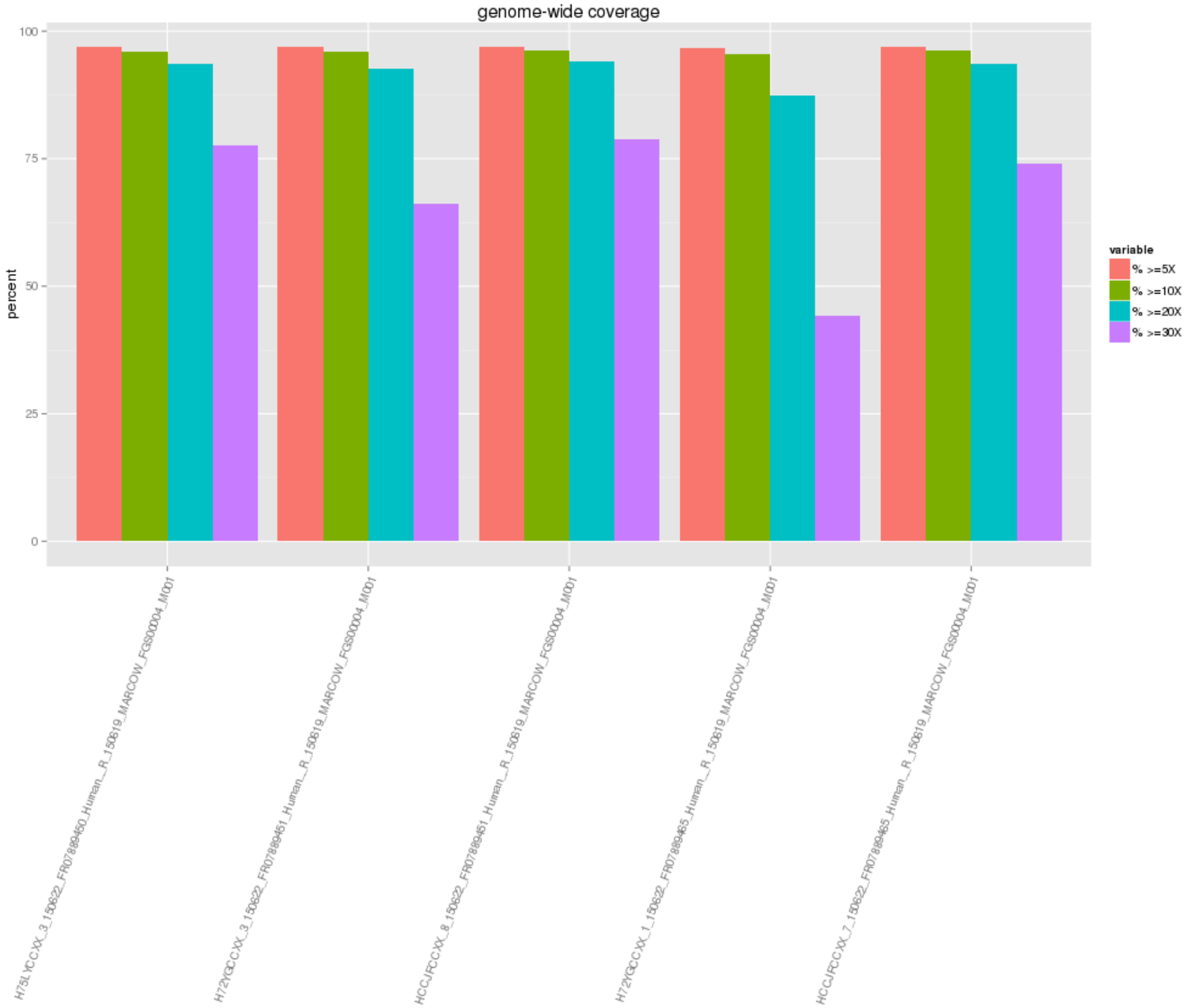

Breakdown of reasons for aligned bases to be excluded
percent
0
10
20
variable
% EXC MAPQ
% EXC DUPE
% EXC UNPAIRED
% EXC BASEQ
% EXC OVERLAP
% EXC CAPPED
50622_FR07889450_Human_R_150819_MARCOW_FGS00004_M001
50622_FR07889451_Human_R_150819_MARCOW_FGS00004_M001
50622_FR07889451_Human_R_150819_MARCOW_FGS00004_M001
50622_FR07889465_Human_R_150819_MARCOW_FGS00004_M001
50622_FR07889465_Human_R_150819_MARCOW_FGS00004_M001
H75LYCCXX_3_1
H72YGCCXX_3_1
HCGJFCCXX_8_1
H72YGCCXX_1_1
HCGJFCCXX_7_1

Coverage, expressed as % of target covered to certain depth

The figure and table below summarise the proportion of the genome that is sequenced above a certain average (e.g. >20X).

> PCT_5X: The fraction of bases that attained at least 5X sequence coverage in post-filtering bases.
>
> PCT_10X: The fraction of bases that attained at least 10X sequence coverage in post-filtering bases.
>
> PCT_20X: The fraction of bases that attained at least 20X sequence coverage in post-filtering bases.
>
> PCT_30X: The fraction of bases that attained at least 30X sequence coverage in post-filtering bases.

	IUS	% >=5X	% >=10X	% >=20X	% >=30X
1	H75LYCCXX_3_150622_FR07889450_Human__R_150619_JOHDOE_FGS_M001	96.89	96.09	93.63	77.51
2	H72YGCCXX_3_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	96.83	95.99	92.60	66.26
3	HCCJFCCXX_8_150622_FR07889451_Human__R_150619_JOHDOE_FGS_M001	96.95	96.19	94.11	78.75
4	H72YGCCXX_1_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	96.64	95.57	87.32	44.29
5	HCCJFCCXX_7_150622_FR07889465_Human__R_150619_JOHDOE_FGS_M001	96.91	96.10	93.61	73.95


genome-wide coverage
percent
100
75
50
25
0
variable
% >=5X
% >=10X
% >=20X
% >=30X
H75LYCCXX_3_150622_FR07889450_Human_R_150619_MARCOW_FGS00004_M001
H72YGCCXX_3_150622_FR07889451_Human_R_150619_MARCOW_FGS00004_M001
HCCJFCCXX_8_150622_FR07889451_Human_R_150619_MARCOW_FGS00004_M001
H72YGCCXX_1_150622_FR07889465_Human_R_150619_MARCOW_FGS00004_M001
HCCJFCCXX_7_150622_FR07889465_Human_R_150619_MARCOW_FGS00004_M001

Methods

Library kits used:

- for Truseq Nano the 'TruSeq Nano DNA Library Prep Kit '
- for PCR-Free the 'TruSeq DNA PCR-Free Library Prep Kit '

Libraries were created as per manufacturer's instructions. One sample was loaded per flow cell lane.

The flow cells were loaded onto an Illumina HiSeq X sequencer and 2x150bp paired-end sequencing was performed. The raw data from the sequencers was converted to FastQ file format using Illumina's bcl2fastq 2.16.0.

To verify data quality, FastQC was run on the FastQ files. Furthermore, the sequences are aligned to b37d5 human reference genome (human.g1k.v37) using Illumina's iSAAC aligner v01.14.11.11 (Raczy et al, 2013; link to whitepaper) to generate BAM files. Additional quality metrics were calculated using Picard WgsMetrics v1.119.

To calculate the mean coverage the following equation was applied as described by the manufacturer.

((MEAN_COVERAGE) x (1 - PCT_EXC_DUPE - PCT_EXC_OVERLAP))/(1 - PCT_EXC_TOTAL)

Further Information

For more details please visit the KCCG Web Page.

Garvan Institute of Medical Research
384 Victoria Street, Darlinghurst, NSW 2010
T: + 61 (0)2 9355 5846 I E: kccgseq@garvan.org.au
http://www.garvan.org.au/kccg