**RDT Capture Evaluation Report**

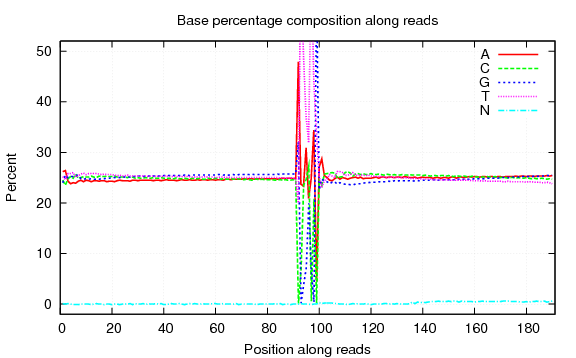
**Chen Dan**

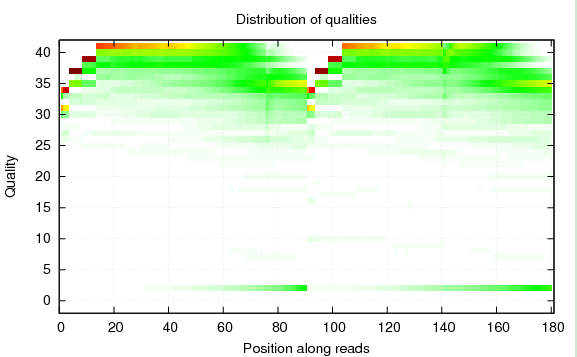
**2011-12-21**

## Basic bioinformatics analysis

## 1.1 QC figures of original data

Figure 1: Base percentage composition and qualities distribution of 10082270





### 1.2 Initially analysis of sequencing data

We take thehuman genome (hg19) chromosome 1~22 and X, Y, M as the reference genome for this project. The genome size of *hg19* is 3,137,161,264, while the effective size is 2,897,310,462 (excluding the N base in the reference). Then BWT-based alignment is carried out for the capture-Sample. We then use total map reads to process the basic data statistics analysis about the relevant items, and after that we use only unique map reads to conduct subsequent analysis.

### 1.2.1 Summary of sequencing data

The dataset is saved as tab-separated values in a text file.

Table 1 Summary of sequencing data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Insert size** | **Read Length** | **Bases(G)** | | **PE map rate (%)** | **Single map rate (%)** | **Mismatch Rate (%)** | **Duplication Rate(%)** | **Mismap Rate(%)** | | **Unmap Rate (%)** | **proper rate** | |
| **Raw Data** | **Filter N&B** |
| 10082270 | 250 | 90 | 0.862 | 0.775 | 85.62 | 6.02 | 0.62 | 1.71 | | 3.58 | 5.05 | | 70.81 | |
| 1000075 | 250 | 90 | 0.996 | 0.896 | 85.84 | 5.81 | 0.6 | 1.19 | | 3.31 | 4.77 | | 71.55 | |
| 10082275 | 250 | 90 | 1.078 | 0.969 | 87.77 | 4.48 | 0.81 | 9.5 | | 3.37 | 4.38 | | 71.20 | |
| 11050458 | 250 | 90 | 1.038 | 0.910 | 88.66 | 4.32 | 0.81 | 0.25 | | 1.50 | 5.52 | | 70.11 | |
| 10082281 | 250 | 90 | 1.435 | 1.216 | 91.34 | 3.63 | 0.68 | 2.00 | | 1.35 | 3.67 | | 80.71 | |
| 10102841 | 250 | 90 | 1.565 | 1.317 | 91.52 | 3.57 | 0.67 | 1.00 | | 1.08 | 3.83 | | 79.26 | |
| 10102843 | 250 | 90 | 0.864 | 0.637 | 90.41 | 3.93 | 0.69 | 2.00 | | 1.58 | 4.09 | | 80.67 | |
| 10102850 | 250 | 90 | 1.406 | 1.180 | 91.85 | 3.36 | 0.74 | 1.94 | | 1.32 | 3.47 | | 81.05 | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Mapped to MHC Data （bp）** | **Mapped to RDT Data (bp)** | **Mapped to REF Data (bp)** | **Target region Rate (%)** | **Total sam number** | **Uniq sam number** | **Uniq sam Rate(%)** |
| 10082270 | 300,569,416 | 300,201,006 | 694,224,360 | 43.24 | 8,724,162 | 6,872,846 | 78.78 |
| 1000075 | 307,040,237 | 306,652,113 | 805,948,311 | 38.05 | 10,081,924 | 8,023,065 | 79.58 |
| 10082275 | 377,403,269 | 376,716,603 | 884,786,535 | 42.58 | 10,924,954 | 8,439,763 | 77.25 |
| 11050458 | 367,537,020 | 366,458,858 | 819,390,641 | 44.72 | 10,217,528 | 7,541,431 | 73.81 |
| 10082281 | 673,769,674 | 673,331,987 | 1,129,616,290 | 59.61 | 13,723,006 | 11,895,394 | 86.68 |
| 10102841 | 680,346,228 | 679,898,718 | 1,213,804,106 | 56.01 | 14,860,114 | 12,721,146 | 85.61 |
| 10102843 | 408,335,528 | 408,076,987 | 593,605,667 | 68.75 | 7,196,884 | 6,221,776 | 86.45 |
| 10102850 | 704,094,816 | 703,682,910 | 1,094,589,898 | 64.29 | 13,291,334 | 11,542,999 | 86.85 |

### 1.2.2 Distribution of per-base sequencing depth

1. For genome re-sequencing, the distribution of per-base sequencing depth should be in accordance with the Poisson distribution theoretically. However, due to the differences between the sample and the reference, the actual distribution might show divergence from the theoretical value. Therefore, this analysis can be used to identify differences between the genome of sample species and the reference. In the case of high genome similarity, the analysis can reveal the actual sequencing depth.

Table 2 Depth distribution and coverage data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 10082270 | 1000075 | 10082275 | 11050458 |
| The mean dep in 300X | 71.99 | 76.38 | 79.75 | 53.86 |
| The mean dep of all above 1 X | 84.02 | 85.88 | 104.10 | 102.00 |
| The Fraction of target covered with at least 4X | 92.21 | 91.90 | 93.72 | 92.27 |
| The Fraction of target covered with at least 10X | 89.52 | 89.33 | 91.77 | 88.02 |
| The Fraction of target covered with at least 20X | 85.79 | 85.89 | 88.44 | 76.60 |
| The Fraction of target covered with at least 30X | 81.13 | 81.75 | 84.00 | 62.27 |
| The Fraction of target covered with at least 300X | 2.18 | 1.55 | 5.00 | 8.64 |
| The Standard Deviation of sample depth | 23.06 | 23.87 | 25.88 | 21.22 |
| The coverage rate of RDT | 94.97 | 94.91 | 96.20 | 95.57 |

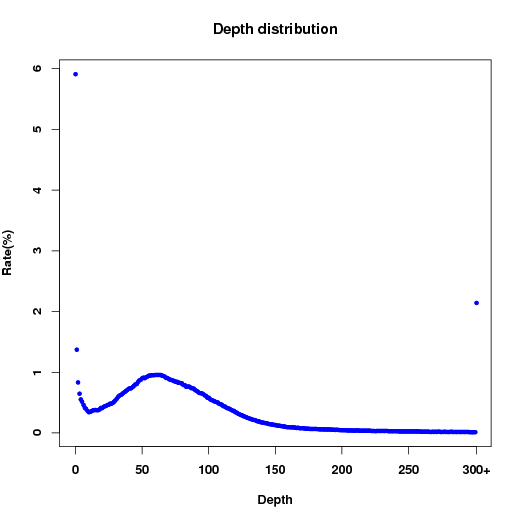
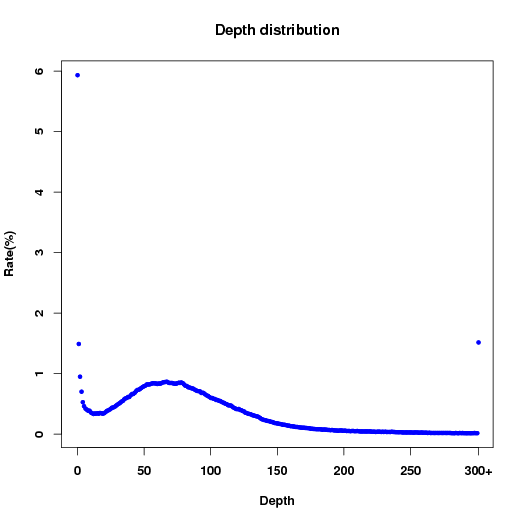
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 10082281 | 10102841 | 10102843 | 10102850 |
| The mean dep in 300X | 171.40 | 174.30 | 108.08 | 134.08 |
| The mean dep of all above 1 X | 189.87 | 190.58 | 115.44 | 196.63 |
| The Fraction of target covered with at least 4X | 91.36 | 92.16 | 90.96 | 91.36 |
| The Fraction of target covered with at least 10X | 88.60 | 89.87 | 87.89 | 88.80 |
| The Fraction of target covered with at least 20X | 85.27 | 87.03 | 83.78 | 85.79 |
| The Fraction of target covered with at least 30X | 82.81 | 84.38 | 80.32 | 83.22 |
| The Fraction of target covered with at least 300X | 1.88 | 1.63 | 0.09 | 19.49 |
| The Standard Deviation of sample depth | 30.33 | 30.46 | 26.33 | 31.12 |
| The coverage rate of RDT | 94.27 | 94.83 | 93.97 | 94.93 |

Table 3: Distribution of uncovered region base number of the samples divided as the twice experiment

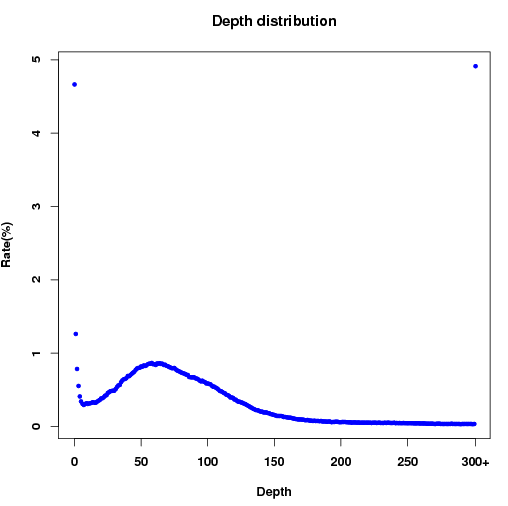
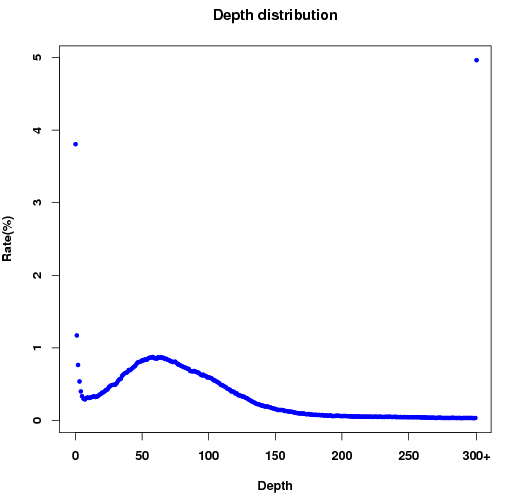
|  |  |  |  |
| --- | --- | --- | --- |
|  | Number | Rate in uncov region (%) | Rate in RDT region (%) |
|
| The common uncovered loci in four samples | 83092 | 23.44 | 2.18 |
| The common uncovered loci in three samples | 67321 | 18.99 | 1.77 |
| The common uncovered loci in two samples | 91859 | 25.91 | 2.41 |
| The common uncovered loci in one sample | 112231 | 31.66 | 2.95 |
| The loci of being in exon region | 91056 | 25.69 | 2.39 |
|  |  |  |  |
|  | Number | Rate in uncov region (%) | Rate in RDT region (%) |
|
| The common uncovered loci in four samples | 132872 | 37.81 | 3.49 |
| The common uncovered loci in three samples | 74246 | 21.13 | 1.95 |
| The common uncovered loci in two samples | 67767 | 19.28 | 1.78 |
| The common uncovered loci in one sample | 76533 | 21.78 | 2.01 |
| The loci of being in exon region | 86684 | 24.67 | 2.28 |

Figure 2: Distribution of per-base sequencing depth curve

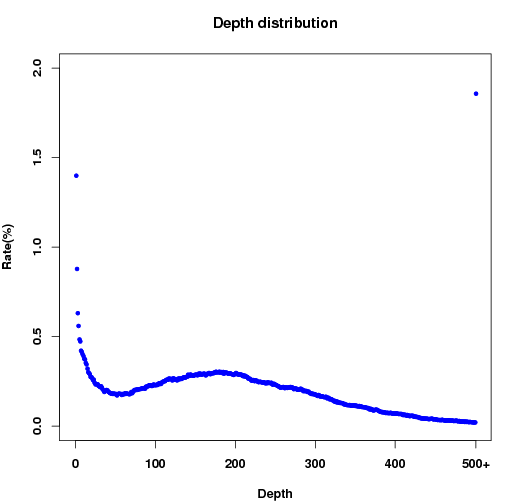
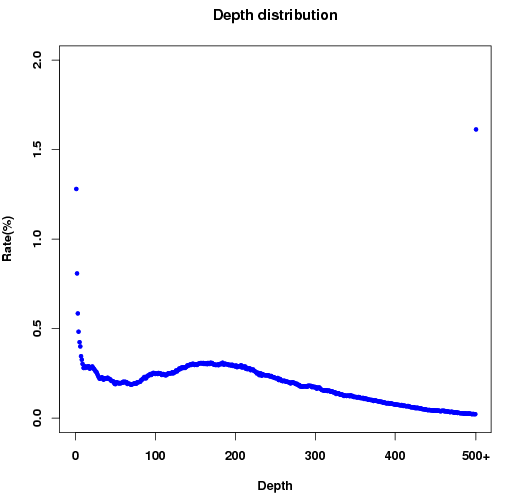
10082270 1000075

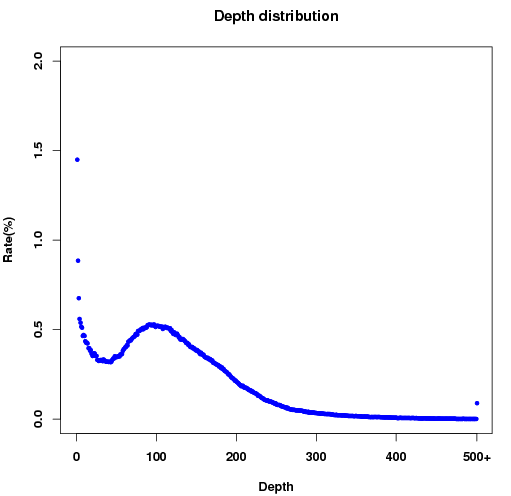
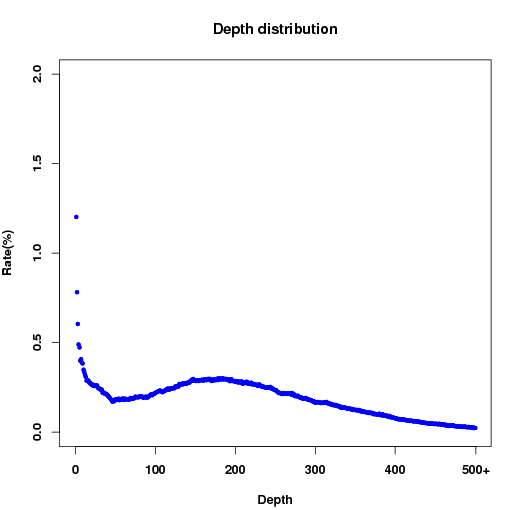
10082275 11050458



10082281 10102841

10102843 10102850

We also add the information of coverage about TNF and MICA gene, which are important for RA

Table 4: Distribution of coverage in the TNF and MICA region

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | TNF | | MICA | |
| cov rate | mean Dep | cov rate | mean Dep |
| 10082270 | 100.00 | 58.54 | 100.00 | 142.08 |
| 1000075 | 100.00 | 56.50 | 100.00 | 134.09 |
| 10082275 | 100.00 | 65.14 | 100.00 | 192.14 |
| 11050458 | 100.00 | 62.28 | 99.90 | 211.07 |
| 10082281 | 100.00 | 162.15 | 100.00 | 247.08 |
| 10102841 | 100.00 | 175.25 | 100.00 | 231.95 |
| 10102843 | 100.00 | 96.45 | 100.00 | 149.86 |
| 10102850 | 100.00 | 180.69 | 100.00 | 231.69 |

Then we still add the coverage distribution about MHC-I and MHC-II gene, details are as follow:

Table 5: Distribution of coverage in the MHC-I and MHC-II region

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MHC-I | | MHC-II | |
| cov rate | mean Dep | Cov rate | mean Dep |
| 10082270 | 100.00 | 197.06 | 91.82 | 65.17 |
| 1000075 | 100.00 | 199.98 | 92.29 | 69.48 |
| 10082275 | 100.00 | 251.65 | 94.82 | 93.00 |
| 11050458 | 100.00 | 275.32 | 96.10 | 91.10 |
| 10082281 | 100.00 | 238.26 | 91.45 | 175.18 |
| 10102841 | 100.00 | 244.13 | 91.98 | 174.58 |
| 10102843 | 100.00 | 138.39 | 85.98 | 97.06 |
| 10102850 | 100.00 | 238.77 | 91.56 | 171.36 |

## 2 Advanced bioinformatics analysis

### 2.1 Assembly of consensus sequences

According to the alignment with reference sequence, under a Bayesian model, the genotype with the highest probability at a locus can be identified for the individual sequencing sample. The consensus sequence of the sample is assembled and saved as CNS format.

### 2.2 SNP detection by comparing captured database with WGS

1. Based on the consensus sequence, the captured SNPs within the RDT region are compared with those of the identified genotype and a fidelity SNP dataset is generated. The dataset is saved as tab-separated values in a text file.
2. RDT capture consensus file filter criteria:

Minimum filter depth=4; maximum filter depth=500; consensus quality >=20; maximum mapping quality >=20.

Table 6: Information of SNP detection by comparing captured database with WGS SNPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Lib | RDT capture SNPs | | WGS  target region SNPs | number of common SNPs | Coverage in RDT SNPs (%) |
| 10082270 | 17430 | 19654 | | 15443 | 88.60 |
| 1000075 | 14897 | 16593 | | 12890 | 86.53 |
| 10082275 | 14766 | 17030 | | 13029 | 88.24 |
| 11050458 | 13916 | 17978 | | 12708 | 91.32 |
| 10082281 | 15658 | 16118 | | 12851 | 82.07 |
| 10102841 | 17617 | 18614 | | 14918 | 84.68 |
| 10102843 | 16576 | 23032 | | 15299 | 92.30 |
| 10102850 | 17311 | 22994 | | 15513 | 89.61 |

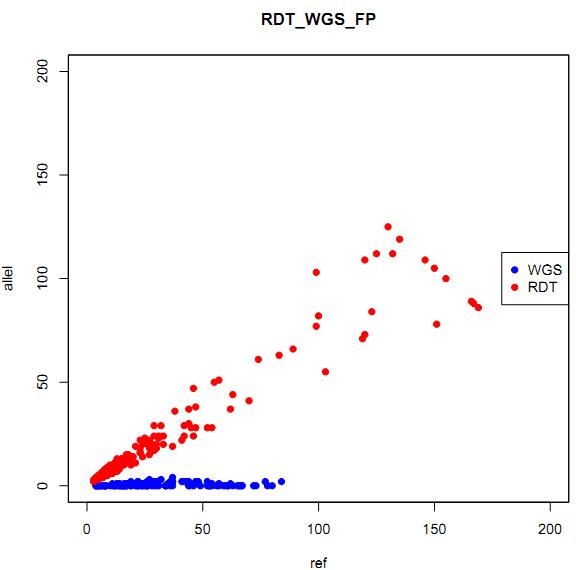
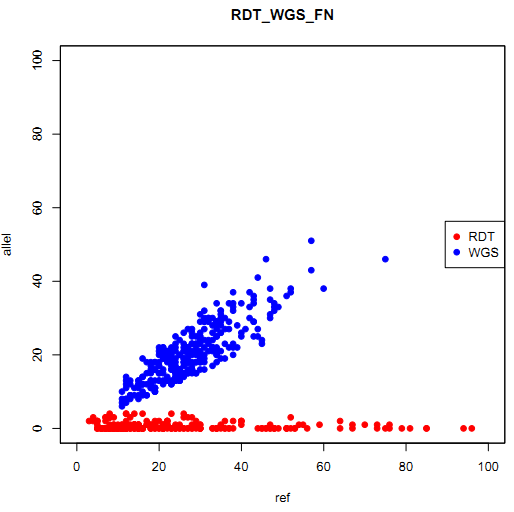
We select respectively the locus fitting our stringent conditions from the WGS of the two samples (11050458 and 10082281) as credible sites, according to which we evaluate the FN and FP of the RDT capture data.

Table6: Information of the SNPs as FN and FP between WGS and RDT

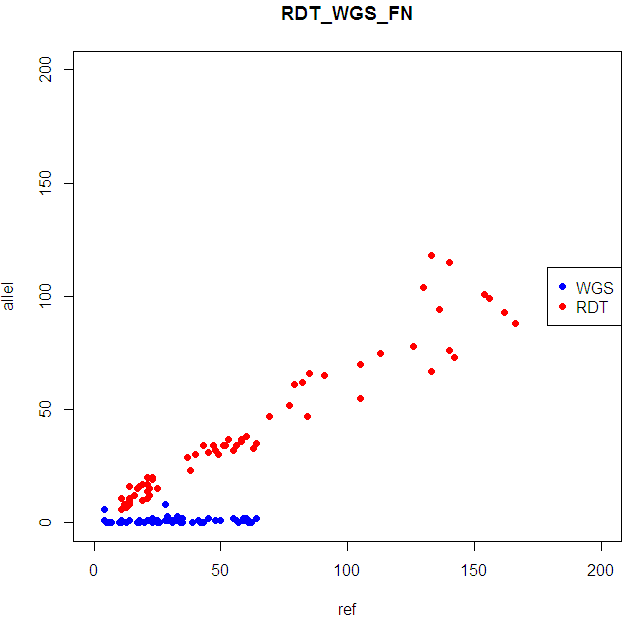
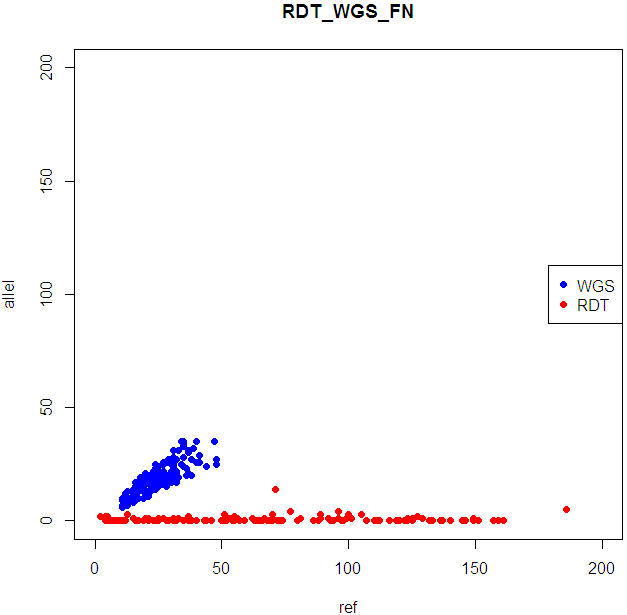
|  |  |  |
| --- | --- | --- |
|  | Number | Rate(%) |
| Select total FN SNPs | 1207 | -- |
| Quality<20 FN SNPs | 313 | 25.93 |
| Depth<4 and depth>500 FN SNPs | 619 | 51.28 |
| Other reasons FN SNPs | 275 | 22.78 |
| Select total FP snp | 250 | -- |
| Quality<20 FP SNPs | 93 | 37.20 |
| Depth<4 or depth>500 FP SNPs | 13 | 5.20 |
| Others reasons FP SNPs | 144 | 57.60 |

|  |  |  |
| --- | --- | --- |
|  | Number | Rate(%) |
| Select total FN SNPs | 728 | -- |
| Quality<20 FN SNPs | 91 | 12.50 |
| Depth<4 and depth>500 FN SNPs | 432 | 59.34 |
| Other reasons FN SNPs | 203 | 27.88 |
| Select total FP snp | 233 | -- |
| Quality<20 FP SNPs | 88 | 37.77 |
| Depth<4 or depth>500 FP SNPs | 49 | 21.03 |
| Others reasons FP SNPs | 96 | 41.20 |

Figure 4: The depth distribution about FN and FP of the sample 11050458



The depth distribution about FN and FP of the sample 10082281



### 2.3 11050458 capture lib500 compared with Genotyping

We classified both the genotyping alleles and the alleles that were called by the capture sequencing of 11050458 into four categories: (1) HOMref. (homozygotes where both alleles are identical to the reference); (2) HOMmut. (homozygote where both alleles differ from the reference); (3) HET ref. (heterozygote where only one allele is identical to the reference); and (4) HETmut. (heterozygote where both alleles differ from the reference and also differ from one another). The number of whole genome sequencing sites that are consistent with capture sequencing at both alleles, at one allele, or that are inconsistent at both alleles were categorized as 2, 1, and 0, respectively.

Because there are not corresponding genotyping alleles in the last 4 samples, so we have omitted this part this time.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Comparison of Capture sequencing and Genotyping alleles** | | | | | | | | |
|  | Allele type | | RDT Capture sequecing SNPs | | | | Total | Consistency (%) |
|  |  |  | HOM ref. | HOM mut. | HET ref. | HET mut. | |  |
| Illumina genotyping | HOM ref. | 2 | 2786 | - | - | - | 2793 | 99.75 |
|  | 1 | - | - | 7 | - |  |  |
|  | 0 | - | 0 | - | 0 |  |  |
| HOM mut. | 2 | - | 742 | - | - | 751 | 98.80 |
|  | 1 | - | - | 5 | 0 |  |  |
|  | 0 | 4 | 0 | 0 | 0 |  |  |
| HET ref. | 2 | - | - | 1357 | - | 1401 | 96.86 |
|  | 1 | 28 | 13 | 3 | 0 |  |  |
|  | 0 | - | 0 | - | 0 |  |  |
| HET mut. | 2 | - | - | - | 0 | 3 | 0 |
|  | 1 | - | 0 | 0 | 0 |  |  |
|  | 0 | 3 | 0 | 0 | - |  |  |
|  | Total |  |  | 4948 |  |  |  | 98.73 |

## Summary

The result reflect some issues which we should pay attention to：

1. specificity

We have modified the result of all the 8 samples, and the target region rate has increased to 50%~70%, some of which are also lower than the predicting 60%.

1. Coverage

For each of the test sample, there is about 4%~5% target region uncovered, and among these locus, about 68% is existed in at least two samples.

(The coverage of the behind 4 samples are a little lower than before for the gel purification we select being 1K~2K,which may lost some data under 1K. )

1. Allele bias

From the statistic result of FN rate, Reference base bias account for a lot parts of the variation site.

1. FP

It seems that, there also exists FP site; some of them may be caused by WGS data’s bias, while some of them may be True FP, which reflect some non-specificity of the primer.

1. Next Plan

We are not able to distinguish the reads whether they are existed in the primer or the target region in our data. So we would consider blocking all the SNPs in the primer sites. Or any better suggestions?