Illumina 9K, 90K conversion to VCF

**Converting Illumina AB into VCF**

Illumina data loaded into T3 database as Ref = A\_allele, Alt = B\_allele
The Illumina 90K data can be combined with similar array data and analyzed with website tools but can not be used in PHG, Beagle, or merged with GBS data because the format is not aligned (strand and orientation) with reference genome

1. use samtools faidx with the IWGSC\_WGA\_v1.0 assembly to get Ref allele

2. compare Ref allele to A\_allele and B\_allele

|  |  |  |
| --- | --- | --- |
| match | Ref and Alt changes | genotypes changes |
| Ref = A\_allele | unchanged | unchanged |
| Ref = B\_allele | Ref = B\_allele, Alt = A\_allele | compliment |
| Ref = comp(A\_allele) | Ref = comp(A\_allele)Alt = comp(B\_allele) | unchanged |
| Ref = comp(B\_allele) | Ref = comp(B\_allele)Alt = comp(A\_allele) | compliment |

Compliment genotypes

|  |  |
| --- | --- |
| original | converted |
| 0/0 | 1/1 |
| 0/1 | 1/0 |
| 1/1 | 0/0 |
| ./. | ./. |

**Accuracy check**

3. compared Infinium 90K and Exome capture 2019 HapMap

To check conversion of the Illumina files, I converted the files using the procedure listed above then compared the output to Exome capture 2019 HapMap protocol. The Exome capture protocol had 200 accessions that are the are genotyped with both the Illumina 90K and Exome capture.

Compare Illumina 90K to Exome capture (200 accessions) by Reference assembly SNP

|  |  |  |
| --- | --- | --- |
| Reference SNP | count markers | accuracy |
| Ref = A\_allele | 3781 | 90% |
| Ref = B\_allele | 4401 | 69% |
| Ref = comp(A\_allele) | 2788 | 91% |
| Ref = comp(B\_allele) | 3112 | 68% |
| Total | 14,082 | 79% |

Compare accuracy by Illumina coding (Polymorphisim and Surrounding Sequence)

|  |  |  |
| --- | --- | --- |
| Illumina SNP | count | accuracy |
| A / (C or G) | 4226 | 78% |
| T / (C or G) | 4152 | 78% |
| A/T ambiguous | 6 | 58% |
| C/G ambiguous | 24 | 75% |

Note: This shows that the accuracy is not dependent on the SNP type. For the two ambiguous cases there are not enough data to make a conclusion.

Compare accuracy by allele frequency

|  |  |  |
| --- | --- | --- |
| Allele |  |  |
| Ref = major  |  | 44% |
| Ref = minor |  | 32% |
| Ref = Comp(minor) |  | 46% |
| Ref = Comp(major) |  | 81% |

Note: These results show that accuracy is not dependent on the allele frequency

Compare accuracy using best blast hit for each chromosome instead of the historical position of the marker.

|  |  |  |
| --- | --- | --- |
| Reference SNP |  | accuracy |
| Ref = A\_allele | 4372 | 85% |
| Ref = B\_allele | 5041 | 71% |
| Ref = comp(A\_allele) | 4360 | 85% |
| Ref = comp(B\_allele) | 4857 | 71% |
| Total | 20824 | 79% |

Compare accuracy usng Jason's Fiedlers marker positions

|  |  |  |
| --- | --- | --- |
| Reference SNP |  | accuracy |
| Ref = A\_allele | 3992 | 90% |
| Ref = B\_allele | 3853 | 70% |
| Ref = comp(A\_allele) | 3084 | 70% |
| Rev = comp(B\_allele) | 2835 | 68% |
| Total | 13,764 | 79% |

Note: These results show that Jason’s modifications did not significantly change the accuracy for the 90K array

https://www.illumina.com/documents/products/technotes/technote\_topbot.pdf