Assessing Allelic Configuration Models in Fixed Ploidy Variant Calling Using R - *Betula*

Jasmin Zohren[1], Igor Kardailsky[2], Kåre Lehmann Nielsen[3], Anika Joecker[2], and Richard Buggs[1]

1) School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK; 2) CLC bio, a QIAGEN company, Silkeborgvej 2, Prismet, 8000 Aarhus C., DK; 3) Department of Chemistry and Bioscience, Sohngårdsholmsvej 49, 9000 Aalborg, DK

Genotyping polyploids

Genotyping of SNP loci in polyploids has always been challenging, but may become easier using high-coverage sequencing. The production of such data in turn requires the development of new methods and software that can be used to analyse it in user friendly software such as the CLC Genomics Workbench (GWB). The "Fixed Ploidy Variant Detection" tool of the CLC GWB already takes higher ploidy levels into account, but an explicit evaluation of the locus configuration is currently not provided. We have developed an algorithm that uses read counts and the average base quality to estimate the most likely allelic configuration at each variant position in polyploid samples.

The data

Solanum tuberosum (potato)
One tetraploid individual
Chromosome 11
50x coverage

Betula (birch)
213 individuals
Three different species
60x coverage

Homo sapiens
Illumina Platinum Genomes
Trio data set, chromosome 19
200x coverage

Tab. 1	Mother	Father	Father	Son	Mother	Son
Chm 19	A_1A_2	B_1B_2	B_1B_2	B_mA_m	A_1A_2	A_mB_m
X chm	X_1X_2	XY	XY	X_mY	X_1X_2	X_mY
Chm 19	$A_1A_2B_1B_2$ 4n (1:1:1:1)		$B_1B_2B_mA_m$ 3n (2:1:1)		$A_1A_2A_mB_m$ 3n (2:1:1)	
X chm	X_1X_2X 3n (1:1:1)		XX_m 2n (1:1)		$X_1X_2X_m$ 2n (2:1)	

Allelic configurations of artificial polyploid humans.
See figure 1 in "Tetraploid humans" section for details.

Tetraploid humans

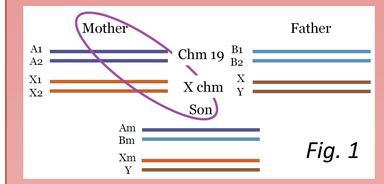
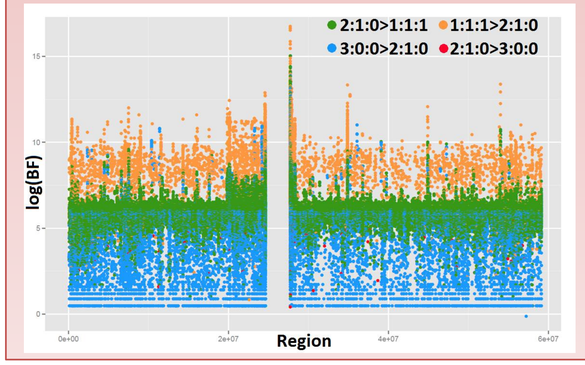


Fig. 1

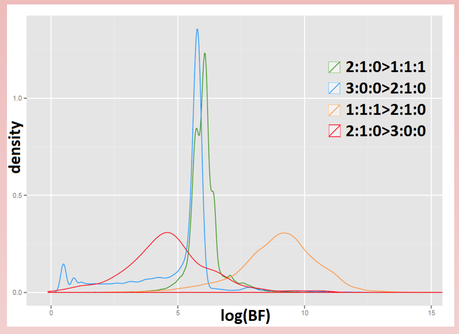
The reads of two individuals from the trio data set are merged, resulting in a polyploid human. The allele dosage differs depending on which individuals are merged and which chromosome is being analysed (see table 1 in "The data"). The data shown here is chromosome 19 of mother and son.

Best Model	#Hits	Second-best models	Mean (BF)	Median (BF)
3:0:0	31,560 (17.7%)	2:1:0	1,211	288
2:1:0	141,271 (79.2%)	3:0:0, 1:1:1*	926	406
1:1:1	5,615 (3.1%)	2:1:0	39,486	8,645
Total	178,446 (100%)			

Distribution of the Bayes factor (BF) for the different model comparisons. A Bayes factor of greater than 5 is considered to be significant.



A total of 178,446 variants were called on the data set of chromosome 19 of mother and son. The "maximum expected ploidy" in the CLC variant caller was set to 3to reflect the actual allele dosage (see table 1 in "The data" section).



Distribution of the different allele dosage models along chromosome 19. More "extreme" models are called with a higher certainty (i.e. Bayes factor) than others. A Bayes factor of greater than 5 is considered to be significant.

How it's done

Input
- Reference sequence
- Sequencing reads

CLC Genomics Workbench
- Quality trimming/Merge overlapping pairs
- Read mapping
- Extract chromosomes/Merge mappings
- Local realignment
- Variant calling
- Extract read counts from every individual

R
- Computation of log-likelihood model
- Most probable allelic configurations
- Bayes factor as a quality measure

Output
- Delimited table
- Visualisations

Next steps

In order to validate the algorithm, more and different combinations of the trio data will be analysed (e.g. father and son and using reads mapping to the X chromosome). This might lead to a refinement of the mathematical model, for example to differentiate between allo- and auto-polyploids. We are also working on a sliding window approach to be able to detect regions with the same allele dosage and make an inference on sequence demography from sudden changes in these regions.
The current version can already be used as a "ploidy detection" tool. Eventually, we intend to implement this as a tool within the CLC Genomics Workbench, probably as an addition to the already existing variant callers.

Funders

The Danish Council for Strategic Research

NATURAL ENVIRONMENT RESEARCH COUNCIL

MARIE CURIE ACTIONS

PEOPLE

Jasmin Zohren

j.zohren@qmul.ac.uk

www.birchgenome.org