Updated phase association and assignment for IRMA v0.7.0 and later: supplemental documentation

Samuel S. Shepard

September 5^{th}, 2019

In the IRMA manuscript [3] we outlined in supplementary (Additional File 10) the variant calling and phase distance measures that were used at that time (v0.6.0). Since IRMA version v0.7.0 we have updated our experimental association distance "EXPENR" to be more accurate. This was done to achieve better phase assignment, which was also added in that version. We review the original material and expand upon it for phase assignment.

Definition of terms

Let D_x be the coverage depth of an assembly at position x, where $1 \leqslant x \leqslant S$. Moreover, let $D_{x,y}$ be the joint coverage depth for any fixed pairs of sites x, y; that is the number of reads that cover both site x and y. Observe that $D_{x,y}$ has properties:

$$D_{x,x} = D_x$$

$$D_{x,y} = 0 \iff x, y \text{ never share a read}$$

$$D_{x,y} \leqslant \min D_x, D_y$$

Let allele $b \in \{A, G, C, T\}$ be at some site x in an assembly s. The count of allele b is denoted $C(b = s_x)$, or rather, the count of allele b at site x in assembly s. We shorten this to $C_{x^\star}$, where the $\star$ reminds us we have selected a particular allele at this site. Let the individual allele frequency be defined as:

$$F(b = s_x) = \frac{C_{x^\star}}{D_x}$$

$$= F_{x^\star}$$

Let $F(b = s_x | s_x \ \& \ s_y)$ be the frequency of allele b at position x given sites x and y co-occur on the same read. Therefore, a conservative estimate of the frequency will be:

$$F_{x^\star}^{\Downarrow} = \min\{F_{x^\star}, F_x(b = s_x | s_x \ \& \ s_y)\}$$

Let $F_{x^\star}^{\Uparrow}$ be the relaxed estimate:

$$F_{x^\star}^{\Uparrow} = \max\{F_{x^\star}, F_x(b = s_x | s_x \ \& \ s_y)\}$$

The joint count of some alleles a and b at sites x and y is denoted $C(a = s_x \ \& \ b = s_y)$ or $C_{x^\star,y^\star}$ for short. This is the co-occurence on the same read. Let the joint frequency be:

$$F_{x^\star,y^\star} = \frac{C_{x^\star,y^\star}}{D_{x,y}}$$

We now have enough definitions to review association measures for minor variants. We had a slightly modified Jaccard distance [1]:

$$J_{x^\star,y^\star} = 1 - \frac{F_{x^\star,y^\star}}{F_{x^\star}^{\Uparrow} + F_{y^\star}^{\Uparrow} - F_{x^\star,y^\star}} \tag{1}$$

a slightly modified, non-log mutual dependency distance [2]:

$$M_{x^\star,y^\star} = 1 - \frac{F_{x^\star,y^\star}^2}{F_{x^\star}^{\Uparrow} \times F_{y^\star}^{\Uparrow}} \tag{2}$$

and our original "experimental association distance":

$$E_{x^\star,y^\star} = 1 - \frac{F_{x^\star,y^\star} \times \min\{F_{x^\star}^{\Downarrow} F_{y^\star}^{\Downarrow}\}}{F_{x^\star}^{\Uparrow} \times F_{y^\star}^{\Uparrow}} \tag{3}$$

Minimums and maximums in the formula render conservative estimates of association. The simple joint probability can also be converted to a distance:

$$1 - 2F_{x^\star,y^\star}$$

Revision of our experimental distance measure

Based on an empirical analysis of known variants utilizing our distance measures, we observed generally good results for the Jaccard distance [data not shown]. Unfortunately, for small sample sizes, Jaccard distances are known not to perform well. To deal with this issue we re-defined our "experimental enrichment" distance to include the Jaccard distance for the majority of cases, except where data are small or for conditions where Jaccard and Mutual Dependency distances might otherwise assume an absolute phasing of variants:

$$E'_{x^\star,y^\star} = \begin{cases} E_{x^\star,y^\star}, & D_{x,y} \leqslant 20 \\ & \text{or } F^2_{x^\star,y^\star} = F^{\Uparrow}_{x^\star} \times F^{\Uparrow}_{y^\star} \\ & \text{or } 2F_{x^\star,y^\star} = F^{\Uparrow}_{x^\star} + F^{\Uparrow}_{y^\star} \\ J_{x^\star,y^\star}, & \text{otherwise} \end{cases} \tag{4}$$

Notice that $F^2_{x^\star,y^\star} = F^{\Uparrow}_{x^\star} \times F^{\Uparrow}_{y^\star}$ when $F_{x^\star,y^\star} = F^{\Uparrow}_{x^\star} = F^{\Uparrow}_{y^\star}$, implying a total phasing of variants as measured by $M_{x^\star,y^\star}$ above. Moreover, $2F_{x^\star,y^\star} = F^{\Uparrow}_{x^\star} + F^{\Uparrow}_{y^\star}$ also when $F_{x^\star,y^\star} = F^{\Uparrow}_{x^\star} = F^{\Uparrow}_{y^\star}$ under Jaccard when variants are fully phased. The original $E_{x^\star,y^\star}$ function is more conservative than both because it scales the joint probability in the numerator by the lowest possible individual frequency of either minor variant.

Phase assignment

Minor variants at the same site can never be in-phase with one another because they cannot co-occur. At the same time, an SNV always co-occurs with itself. This is pictured in the heat-maps produced at the end of an IRMA assembly, with self-phasing on the diagonal. If fewer than 2 variants are called, no phasing may be calculated. However, if 2 or more minor variants are called, IRMA will try to assign phases using the same distance matrix it uses to construct heat-maps for ***qualitative phase analysis***. Specifically, we use the $E'_{x^\star,y^\star}$ function to generate a revised "EXPENR" distance matrix for ***quantitative phase assignment***.

Phase assignment is performed by creating a tree from the distance matrix and cutting the branches at a specified height to provide phase groups. The default height was selected empirically on known mixture data to render 95% sensitivity and specificity [data not shown]. IRMA uses the `R` statistical program's `hclust` function for single-linkage agglomerative clustering followed by the `cutree` function to slice minor variants into phase groups. If minor variants are phased, they will belong to the same group and have the same phase number. If variants are not in-phase, they will have different phase numbers. If no variants are phased, each minor variant will have its own unique phase number. Phase numbers are added to the variants tables written at the end of an IRMA assembly.

Disclaimer

The findings and conclusions in this document are those of the author and do not necessarily represent the views of the Centers for Disease Control and Prevention or the Agency for Toxic Substances and Disease Registry.

References

[1] Dipak L Chaudhari, Om P Damani, and Srivatsan Laxman. Lexical co-occurrence, statistical significance, and word association. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1058–1068. Association for Computational Linguistics, 2011.

[2] Pavel Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–61. Citeseer, 2008.

[3] Samuel S Shepard, Sarah Meno, Justin Bahl, Malania M Wilson, John Barnes, and Elizabeth Neuhaus. Viral deep sequencing needs an adaptive approach: Irma, the iterative refinement meta-assembler. *BMC Genomics*, 17:708, 09 2016.