### SUMMARY

Current methods to find variation based on a reference genome rely on two basic steps: the first one is to map the sequenced reads to the reference genome allowing some number of differences; the second step consists in scanning the alignments to find variable positions applying several filters, e.g. discard ambiguously mapped reads and eliminate variation that is inconsistent with population frequency data. This results in repetitive regions, such as transposons, being very difficult to analyze and de novo variation being exceptionally challenging to identify.

We have developed a method to overcome these challenges and precisely localize variation by exclusively taking into account perfect matches between the reads and unique strings found in the reference genome (COIN-strings). This method is called “Context Dependent Individualization of Nucleotides and Virtual Genomic Hybridization (COIN-VGH)."

### METHODS

#### COIN-VGH PIPELINE

1) CS’s are hybridized against the whole sequencing project database. The number of reads containing each CS is counted to construct the CS Landscape.

2) A variation will produce a sharp reduction in the coverage along the CS landscape.

3) The reads containing the flanking CS’s will be locally aligned to pinpoint the specific base pair and type of variation found. (Reyes, Gómez-Romero, 2011)

### MATHEMATICAL FRAMEWORK

Coverage changes are measured using

\[ F_n = \frac{X_{a+n} - X_n}{\max(X_{a+n}, X_n) \cdot 1} \]

where \( X_n \) equals the number of reads mapped to the CS \( n \).

Genotype likelihood

\[ L(g_j) = \prod_i \left( \frac{(m - g_i) e + g_i (1 - e_j)}{m} \right) \cdot \left( \frac{(m - g_i) (1 - e_j) + g_i e_j}{m} \right) \]

where \( g_i \) refers to the number of reference alleles in the genome of individual \( x \), \( x \in \{ \text{child, father, mother} \} \), \( m \) equals the ploidy. At a specific position there are \( k \) reads piled up, of those are identical to the reference and \( k - i \) are different, \( i \) is the error rate of read \( j \).

**De novo** confidence

\[ D_n = -2 \log \left[ \frac{\max(g_j, g_m) \in \{ L(g_j) L(g_m) \}}{\max L(g_j) \cdot \max L(g_m)} \right] \]

where the numerator refers to the maximum product of the likelihoods of all possible mendelian patterns and the denominator refers to the product of the maximum likelihood genotypes in each individual. This framework was previously implemented by SAMtools. (Li, 2011)

### RESULTS

**PROOF OF CONCEPT – VENTER GENOME chrX**

The COIN-Strings along the chrX (excluding the pseudoautosomal regions) were obtained. The whole Venter genome sequencing project database was analyzed by COIN-VGH.

**DIFFERENT VARIATION TYPES: COIN-VGH LANDSCAPE**

**DE NOVO VARIATION**

A family trio was sequenced (30X average coverage). The child sequencing database was analyzed by COIN-VGH to identify the variable regions. The flanking CS’s were hybridized against the parental reads. A genotype was assigned for each individual. Possible de novo SNV’s were identified. These candidates were filtered by coverage (>20X) and purity (homozygous sites must have no alternative reads).

**CONCLUSION AND FUTURE WORK**

- Most of the SNV’s identified follow a mendelian inheritance pattern validating the COIN-VGH strategy.
- De novo variants can be precisely localized using the COIN-VGH strategy.
- Future work needs to be done to identify other kinds of de novo variants and experimental validation of de novo SNV’s found remains to be done.

### REFERENCES