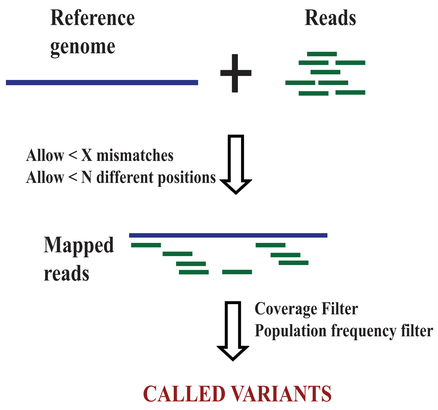
COIN-VGH: Sensitive and Specific Virtual Genomic Hybridization to pinpoint de novo variations

Laura Gómez-Romero[1,2] José Reyes[1] Kim Palacios-Flores[1,2] Ximena Ibarra-Soria[1] Delfino García[1] Margareta Boege[1] Guillermo Dávila[1] Margarita Flores[1]
Michael Schatz[2] and Rafael Palacios[1]
1 Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico
2 Cold Spring Harbor Laboratory, Cold Spring Harbor, USA

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)



String	Recurrences number
C	~750 millions
TGGCACC	318,776
GGGCCTGGCACCAGGGA	1 (COIN STRING)

CS length	% exon	% total
15 nt	29.47	12.21
25nt	90.04	83.83
50nt	92.72	92.32

Every COIN-string unambiguously localizes all the nucleotides it contains

REFERENCES

- Reyes J, Gómez-Romero L, *et.al. Context-dependent individualization of nucleotides and virtual genomic hybridization allow the precise location of human SNPs*. PNAS. August 2011.
- Levy S, et al. *The diploid genome sequence of an individual human*. PLoS Biol. October 2007.
- DePristo M, *et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data*. Nature Genetics. May 2011.
- Heng Li. *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics. September 2011.

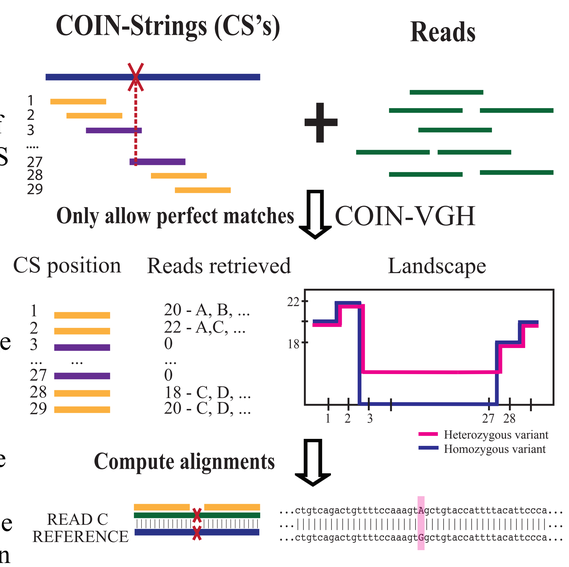
METHODS

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

$$F_n = \frac{X_{n+1} - X_n}{\max(X_{n+1}, X_n, 1)}$$

where X_n equals the number of reads mapped to the CS n

COIN	#reads	F_n
100	22	-0.09
101	20	-1
102	0	0
103	0	1
104	25	

Genotype likelihood

$$L(g_x) = \prod_{j=1}^{l}\left[\frac{(m-g_x)\varepsilon_j + g_x(1-\varepsilon_j)}{m}\right]\prod_{j=l+1}^{k}\left[\frac{(m-g_x)(1-\varepsilon_j) + g_x\varepsilon_j}{m}\right]$$

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

De novo confidence

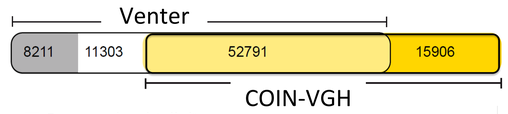
$$D_n = -2\log\left[\frac{\max(g_c, g_f, g_m) \in \{L(g_c)L(g_f)L(g_m)\}}{\max L(g_c) \cdot \max L(g_f) \cdot \max L(g_m)}\right]$$

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

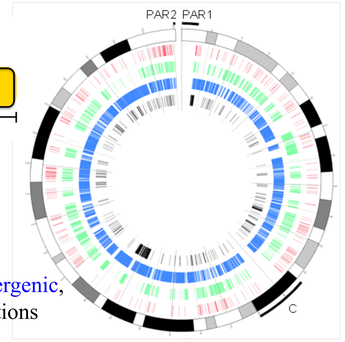
PROOF OF CONCEPT – VENTER GENOME chrX

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

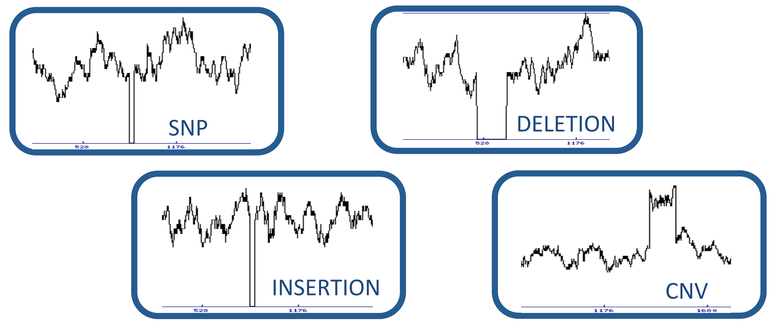


Notably COIN-VGH reported nearly 0 biallelic SNV's as expected.

SNV'S found in: exons, introns, intergenic, segmental duplications



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)

Mendelian	*De novo* SNV's
1,762,973	464

NR – Non reference, R- Reference

DE NOVO SNV's (TOTAL 464)			
PARENT A	PARENT B	CHILD	FREQUENCY
R/R	NR/NR	NR/NR	255
R/R	NR/R	NR/NR	169
R/R	R/R	NR/R	39
NR1/NR1	NR2/NR2	NR1/NR1	1

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.