Hereditary Genetics: Current Research

Román-Ponce SI, et al., Hereditary Genet 2016, 5:2
DOI: 10.4172/2161-1041.1000170

Research Article Open Access

Research Note: Evaluation of Resequencing Technologies Parameters for CNV Genotyping

Sergio Ivan Román-Ponce[1,2,3]*, **Alessandro Bagnato**[2] **and Theo Meuwissen**[3]

[1]*Centro Nacional de Investigación en Fisiología y Mejoramiento Animal, Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias, México*

[2]*Università degli Studi di Milano. Dipartimento di Scienze e Tecnologie Veterinarie per la Sicurezza Alimentare, Via Celoria 10. 20133 Milano, Italia*

[3]*Department of Animal and Aqua cultural Sciences, Norwegian University of Life Sciences, Norway*

***Corresponding author:** Dr. Sergio Ivan Roman Ponce, Centro Nacional de Investigación en Fisiología y Mejoramiento Animal, Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias, México, Tel: 01 800 088 22 22 0 (55) 38 71 87 00; E-mail: roman.sergio@inifap.gob.mx

Rec Date: Mar 16, 2016; **Acc Date:** Jul 15, 2016; **Pub Date: Jul 17**, 2016

Abstract

Whole genome (re)sequencing provides new opportunities to discover Copy Number Variation (CNV) on the genome. Due to the continuous reduction in sequencing costs, it has become as the principal methodology to detect CNV in livestock. One parameter that increases the genotyping cost is the depth of the coverage during sequencing. The main aim of this note was to assess the variation on CNV identification with different depth coverage and read-length on genome sequencing. The results point out that sequences coming from short read-length require less depth coverage than those obtained with long read-length. In addition, small CNV require deeper coverage to be detected. These results can reduce the discovering and genotyping costs since sequencing technologies with short read-lengths are often less costly. Finally, a general formula was derived to optimize the sequencing costs.

Keywords: Copy number variation; Depth of coverage; Livestock; Read-length

Introduction

Copy Number Variation (CNV) represent a significant source of genetic diversity in mammals covering ~12% of the genome [1], and it has been shown to be associated with phenotypes (diseases/traits) in humans [2]. Next-Generation Sequencing (NGS) technology allows for whole genome (re)sequencing at very low costs per sequence and provides a wealth of information to tackle genetic problems, such as the identification of the molecular basis of complex traits that are difficult to study with conventional approaches [3]. To discover (detect, validate and characterize) or genotype CNV on whole genome sequences, the array Comparative Genomic Hybridization (aCGH) has been so far the most used technique. In aCGH experiments genomic DNA samples are co-hybridized on the same oligonucleotide array and the genomic variation differences from the reference sample lead to CNV detection [4].

Currently, some studies that identified CNV using aCGH on cattle [5,6], chicken [7], swine [8] and goat [9] are available. The sequencing effort and its cost represent an important limit to the identification of CNV in livestock populations. One of the parameters that deeply affect the genotyping costs is the coverage of the sequencing. The main aim of this note is to assess the effects of depth of coverage (X) and read-length of the sequencer (RL) on the accuracy of the estimate of the number of copies present in a CNV. All these parameters intend to represent the most common resequencing technologies available.

First of all, we need to know the number of reads (Nr) of the sequences, which is calculated as,

$$N_r = \frac{\mathbf{X} * \mathbf{L_g}}{\mathbf{R_L}}$$

Where Lg is the genome length and RL is the read-length of the sequencer.

The number of times that a read is within a CNV (K) is a function of the number of tandem repeats (copies) within the CNV (n), of the size of each copy of the CNV (S), of the RL and of the Lg and it is calculated as follows:

$$E(K) = N_r * \mathbf{n} * \left(\frac{\mathbf{S} - \mathbf{R_L}}{\mathbf{L_g}}\right) (1)$$

Assuming a Poisson distribution of K(X, Klambauer), the variance of K is

$$\mathbf{V(K)} = \mathbf{E(K)} = \mathbf{n} * \frac{X}{\mathbf{R_L}} * \left(\ \mathbf{S} - \mathbf{R_L}\right) (2)$$

Finally, the coefficient of variation of the number of counts (CV) is:

$$CV(K) = \frac{\sqrt{V(K)}}{E(K)} = \left[\sqrt{n * \frac{X}{\mathbf{R_L}} * \left(\mathbf{S} - \mathbf{R_L}\right)}\right]^{-1} (3)$$... which it is used as a measure of the accuracy of the estimates of K.

Input parameters used to assess the accuracy of K were: the length of the bovine genome (Lg = 2,344 megabase) as reported previously [10], the CNV size (S = from 1 to 200 kb) according to the results of Fadista [5], the read-length (RL=30, 90, 150 and 300bp) and the depth of coverage of the sequence (X=10, 20 and 30).

To evaluate the number of times that random fragments with size of read-length (RL= 30, 90, 150 and 300 bp) were inside of one CNV, three different sizes of the CNV were considered: 1.6, 105.5 and 220.1 Kb per copy. The numbers of fragments were extracted randomly *in silico* from the bovine genome sequence and correspond to the number of reads of the genome. Proportions of fragments inside of a CNV [E] were estimated as follows:

$$\mathbf{E}\ (F_{CNV}) = \frac{K}{N_r}\ (4);$$

where represents the number of counts of read fragments inside of a CNV and Nr is the number of reads.

The coefficient of variation of K is a function of the read-length and of the coverage depth. The CV(K) decreases with shorter reads and deeper coverage in the sequence as shown in the Figure 1.

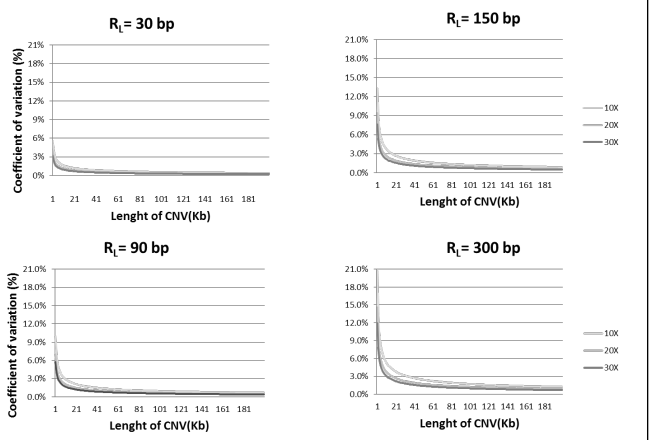

Figure 1: Coefficient of variation (%) estimated in the detection of Copy Number Variation through whole genome sequencing with four Read-Length (RL) and three different depth of coverage (X).

Additionally, when the CNV length increases the coefficient of variation decreases, independently from the depth of the coverage and from the read-length here tested. The number of fragments included in a CNV extracted *in silico* marginally differs from the prediction done by formula [1] (Table 1).

The *in silico* experiment was repeated and the results did not change because the read-length was a constant. The proportion of reads inside of a large (220.1 kb), a medium (105.5 kb) or a small (1.6 kb) CNV were the same for 10X (0.001%), 20X (0.075%) and 30X (0.155%), which shows that depth coverage did not affect the expected copy number estimation of the genotyping, only the accuracy of this estimate.

In cattle, the average size of CNV is 72.3 kb, with a median of 16.7 kb (Min= 1.7 kb; Max= 2,031 kb) [5]. The detection and genotyping of CNV by sequencing depends on the read-length of the sequencer and the size of the CNV. Accounting for these parameters is necessary to determinate the required depth of coverage in order to minimize the cost of genotyping on whole or target (re)sequencing. If whole genome (re)sequencing is used, a deep coverage is recommended to permit the accurate genotyping of also the smallest CNV. However, when a certain region is sequenced to detect one or several CNV(s), the formula [1] can be used to optimize the depth coverage and increase the accuracy of this CNV genotyping; Its application is not restrictive for cattle, can also be used in other organisms where the state of knowledge has not advanced sufficiently in order to optimize the economic effort.

RLa (bp)	Coverage (X)	Nrb (n)	c 1.6 kb, 105.5 Kb, 220.1 Kb			d 1.6 kb, 105.5 Kb, 220.1 Kb		
30	10	47,139,530	528	35,096	73,463	523.3	35,156.7	73,356.7
30	20	94,279,060	1,036	70,387	147,046	1046.7	70,313.3	146,713.3
30	30	141,418,590	1,585	105,490	220,616	1570.0	105,470.0	220,070.0
90	10	15,713,180	166	11,781	24,192	167.8	11,712.2	24,445.6
90	20	31,426,360	317	23,555	48,735	335.6	23,424.4	48,891.1
90	30	47,139,540	484	35,386	73,085	503.3	35,136.7	73,336.7
150	10	9,427,910	121	6,902	14,638	96.7	7,023.3	14,663.3
150	20	18,855,820	221	12,887	29,037	193.3	14,046.7	29,326.7
150	30	28,283,730	336	20,842	43,743	290.0	21,070.0	43,990.0
300	10	4,713,950	50	3,533	7,255	43.3	3,506.7	7,326.7
300	20	9,427,900	95	7,072	14,456	86.7	7,013.3	14,653.3
300	30	14,141,850	144	10,563	21,859	130.0	10,520.0	21,980.0

Note: [a]RL= Read-Length; [b]Nr=Number of fragments; [C] Number of fragments inside CNV *in silico*; [d] Number of fragments inside CNV by formula.

Table 1: Estimated number of copies presented in one copy number variations (CNV) varying the read-length, depth coverage and the size of CNV.

An inherent problem of NGS data the considerable read-mapping ambiguity [11]. Several methods to detect CNV are based on read depths which assume Poisson distribution. Recently, several completely sequenced genomes were examined, and the Poisson distribution assumption was violated by some NGS technologies [12]. Despite this, in this study the results show that sequences obtained from shorter read-length require less depth coverage, a deeper coverage is required when small CNV are searched. Based on this conclusion, the advantage for the scientific community is that technologies with shorter read-length tend also to be less costly.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 222664. ("Quantomics").

Disclaimer

This Publication reflects only the author's views and the European Community is not liable for any use that may be made of the information contained herein.

References

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454.
2. Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61: 437-455.
3. Hobert O (2010) The impact of whole genome sequencing on model system genetics: Get ready for the ride. Genetics 184: 317-319.
4. Shinawi M, Cheung SW (2008) The array CGH and its clinical applications. Drug Discov Today 13: 760-770.
5. Fadista J, Thomsen B, Holm L, Bendixen C (2010) Copy number variation in the bovine genome. BMC Genomics 11: 284.
6. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, et al. (2010) Analysis of copy number variations among diverse cattle breeds. Genome Res 20: 693-703.
7. Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N (2010) An initial map of chromosomal segmental copy number variations in the chicken. BMC Genomics 11: 351.
8. Fadista J, Nygaard M, Holm L, Thomsen B, Bendixen C (2008) A snapshot of CNVs in the pig genome. PloS One 3: e3916.
9. Fontanesi L, Luigi Martelli P, Beretti F, Riggio V, Dall'olio S, et al. (2010) An initial comparative map of copy number variations in the goat (Capra hircus) genome. BMC Genomics 11: 639.
10. Snelling WM, Chiu R, Schein JE, Hobbs M, Abbey CA, et al. (2007) A physical map of the bovine genome. Genome Biol 8: R165.
11. Palmieri N, Schlotterer C (2009) Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. PLoS ONE 4: e6323.
12. Miller CA, Hampton O, Coarfa C, Milosavljevic A (2001) Read-depth: A parallel R package for detecting copy number alterations from short sequencing reads. PLoS ONE 6: e16327.