Characterization of deletions and duplications involved in Deldupemia

Pan Zhang

5/26/2020

1. Trained colleagues summary

Problematic probes: CNSL_probe_5, low sequencing depth.

Deldupemia is an autosomal recessive disease caused by mutations in the CNSL gene. To detect deletions and duplications located at CNSL region, we are developing an NGS-based assay. This summary is about the results of retrospective analysis of 10,000 samples spanning different ethnicities.

Among 477 samples with mutation in CNSL region, 236 samples are from ethnicity A, 158 samples are from ethnicity B, and 83 sample are belong to ethnicity C. The frequency of mutation in each ethnicity is showed in Table 1. More duplication are identified deletion for ethnicity B and C, while for ethnicity A, they are similar. And the overall mutation frequency of ethnicity C is lowest, which may indicate less people of ethnicity C suffering from Deldupemia.

An interest finding is that 100% of the sample from ethnicity A with mutation have one mutation, 74.68% of the sample from ethnicity B with mutation have one mutation, while only 14.46% of the sample with mutation from ethnicity C have one mutation. The average mutation of a carrier from ethnicity C is 2.48. This difference may suggest different mechanisms for mutation generation in different ethnicities.

Table 1. A summary of copy number variation frequency for different ethnicity

ethnicity	duplication (%)	deletion (%)	mutation (%)
A	2.49	2.25	4.74
B	5.19	1.02	6.21
C	2.84	0.57	3.41

A total of 134 breakpoints were identified for depulication, including 3 breakpoints for ethnicity A, 33 breakpoints for ethnicity B, and 96 breakpoints for ethnicity C. In term of breakpoint for deletions, 2, 2, 3 breakpoints are identified for ethnicity A, B, C, respectively. Based on Table 2, the breakpoints for ethnicity A and deletion are stable, while the breakpoint for duplicaton in ethnicity B and C is variable (Table 2).

Table 2. Breakpoints with top frequency (> 0.05) for different ethnicity

Type	ethnicity	5' breakpoint	3' breakpoint	count	Frequency
duplication	A	CNSL_probe_27	CNSL_probe_34	59	0.476

	A	CNSL_probe_32	CNSL_probe_38	59	0.476
	B	CNSL_probe_20	CNSL_probe_40	41	0.234
	B	CNSL_probe_24	CNSL_probe_40	35	0.2
	B	CNSL_probe_37	CNSL_probe_40	14	0.08
	B	CNSL_probe_27	CNSL_probe_40	13	0.074
	C	CNSL_probe_37	CNSL_probe_40	16	0.085
	C	CNSL_probe_10	CNSL_probe_17	11	0.059
deletion	A	CNSL_probe_32	CNSL_probe_38	57	0.509
	A	CNSL_probe_27	CNSL_probe_34	55	0.491
	B	CNSL_probe_20	CNSL_probe_40	18	0.692
	B	CNSL_probe_24	CNSL_probe_40	8	0.308
	C	CNSL_probe_10	CNSL_probe_40	10	0.556
	C	CNSL_probe_10	CNSL_probe_22	4	0.222
	C	CNSL_probe_24	CNSL_probe_40	4	0.222

Another interest finding is that breakpoints correspond with ethnicity. Breakpoints CNSL_probe_27 - CNSL_probe_34 and CNSL_probe_32 - CNSL_probe_38 only appears in ethnicity A, CNSL_probe_20 - CNSL_probe_40 only appears in ethnicity B, and CNSL_probe_10 - CNSL_probe_17 is unique to ethnicity C. In addition, most of the breakpoints with high frequency have been reported in literature.

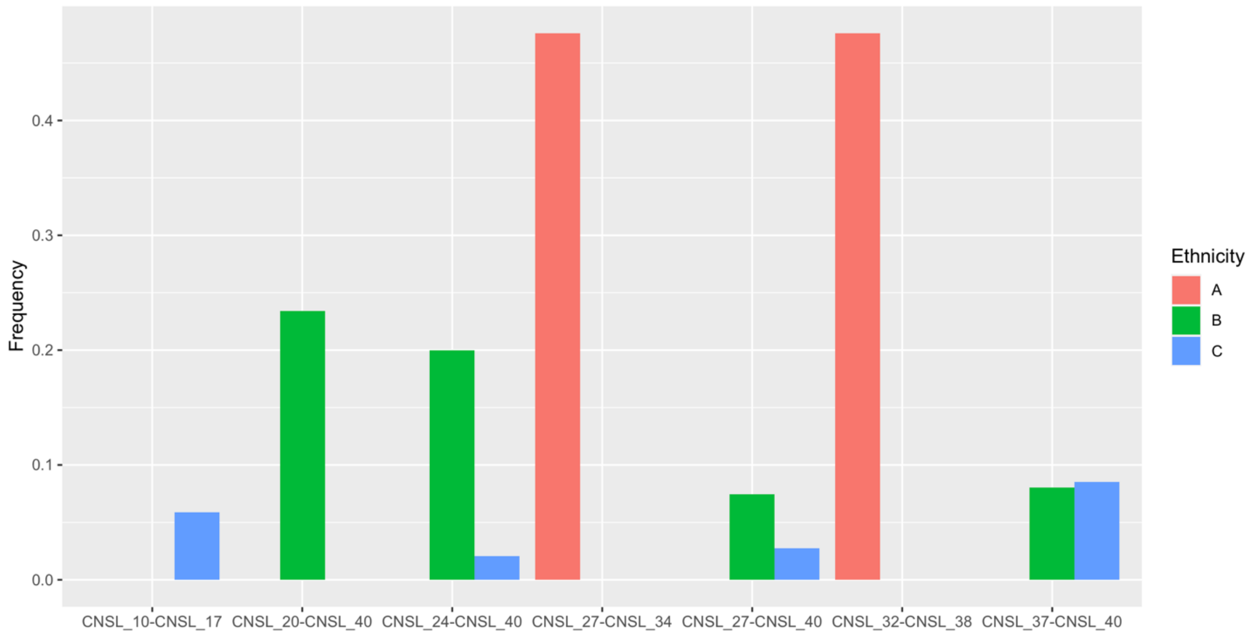

Fig. 1 The distribution of breakpoints among different ethnicity for duplication.

This finding can help predict the ethnicity of a hypothetical set of unlabeled samples. First, we need to predict copy number for each sample. If some sample have duplication or deletion, then we can label some sample through the ethnicity-specific breakpoints. Then the ethnicity with large propotion should the right ethnicity. Another method is to add more labled samples across possible ethnicity, then all samples are clustered based on the mutation of sequencing depth information using unsupervised learning method. The unlabled samples can be labled based on the labled samples and cluster information.

In summary, the mutation frequency and mutation number for each genetic carrier across different ethnicity is different. And the breakpoints correspond with ethnicity, which may help predict the ethnicity of samples.

2. General audience summary

Deldupemia is an autosomal recessive disease caused by mutations in the CNSL gene. The probability for a people to be a genetic carrier is about 6%. This NGS-based assay is developed for detecting the mutation in Deldupemia, which can identify the risk of couples to pass down this inherited disease.

3. Algorithm for finding deletions and duplications

Step 1: filter the prob with less than 10 depth in at least 10 sample (CNSL_probe_5 is filtered in this case)

Step 2: fit the depth to a statistical model descripted on the paper named "Development and validation of a 36-gene sequencing assay for hereditary cancer risk assessment". In brief, the likelihood of observing a given number of mapped reads di,j at a given genomic position i for sample j with copy number ci,j is modeled as following:

$$p(d_{i,j} \mid c_{i,j}) = \mathrm{NegBinom}(d_{i,j} \mid \mu = c_{i,j}\, \mu_i \mu_j , r = r_i)$$

where μi is the average depth for that targeted location across samples, μ_j is the average depth for that particular sample across targeted positions, ri is the dispersion parameter.

Step 3: Assign the copy number with largest probability to each sequencing depth.

Step 4: A deletion or duplication is any contiguous stretch of at least four well behaved probes that have copy number of ~1 or ~3, respectively.