Characterization of deletions and duplications involved in Deldupemia

Goals of this exercise:

1. Give you a sense of whether you enjoy the type of work we do.
2. Give us a sense of your current skills in the area.

Background:

You are developing an NGS-based assay for a recently characterized disorder called Deldupemia, an autosomal recessive disease caused by mutations in the CNSL gene. Due to high sequence homology in the CNSL region, deletions and duplications are common (thus, rather than all patients having two copies of DNA across the region, some have one, and others have three). The deletion/duplication breakpoints can vary from sample to sample (see table further below), and it has been hypothesized that the breakpoints correspond with ethnicity.

With help from someone on the molecular biology team, you developed probes for the CNSL region and performed a retrospective analysis of 10,000 Myriad Women's Health samples spanning different ethnicities. The data in "cnsl_data.csv.gz" catalogs the depth of NGS reads at 100 different hybrid capture probe locations, 50 in the CNSL region, and 50 outside of the region. The depth at a probe is expected to be linearly proportional to the copy number of the DNA at that site (i.e., having three CNSL copies should give roughly 3x the depth as having one copy).

Assumptions:

- Probes in the "nonCNSL" region are expected to have CN=2.
- A deletion or duplication is any contiguous stretch of at least four well behaved probes that have copy number of ~1 or ~3, respectively.
- Due to variability of extraction efficiency in the lab and error in the quantification of DNA libraries, each sample has a slightly different average NGS read depth across all probes.
- Each probe captures DNA with different efficiency relative to other probes, but you can assume that a single probe is equally efficient across all samples.
- The breakpoint positions are known in the literature and correspond to the following probe locations:

Del/dup index	5' breakpoint	3' breakpoint
1	CNSL_probe_32	CNSL_probe_38
2	CNSL_probe_27	CNSL_probe_34
3	CNSL_probe_20	CNSL_probe_40
4	CNSL_probe_10	CNSL_probe_40

Please do the following:

- Write code to characterize the deletion and duplication frequencies and breakpoint positions on a per-ethnicity basis. Your algorithm for finding deletions and duplications may use hardcoded breakpoints based on the table above, or be general. Please return the code you wrote (any language is fine; plotting code not required) to us as a text file.
- Write a summary of your findings containing:
    - At least one paragraph of text (no more than two pages total of text, please) for an audience of scientifically trained colleagues at Myriad Women's Health.
    - Address whatever you think is relevant from your analyses, plus the following:
        - Help the lab identify any problematic probes that may need redesigning.
        - Describe (i.e., no coding required) in just a few sentences how you might predict the ethnicity of a hypothetical set of unlabeled samples where all have the same ethnicity but the ethnicity is not known to you a priori.
    - At least one figure (but not more than four).
    - A ≤4 sentence paragraph (please call it the "General audience summary") that could summarize the goal of the project and its key findings for colleagues in the sales and marketing teams (assuming minimal technical knowledge).

Don't hesitate to email gould@counsyl.com if you hit a wall or would like assistance/clarification. We've tried to leave this pretty open-ended so that you're free to explore and interpret the data as you see fit, but we also attempted for this analysis to consume ten hours or less of your time. So, if it is ballooning well beyond that, please don't hesitate to reach out so we can work with you.