BioSpotlight

Minimal Set, Maximally Informative

Whole-genome studies are performed regularly, resulting in datasets composed of thousands of individuals genotyped with hundreds of thousands of markers. This information is valuable in identifying disease-associated allelic variants, as well as discerning patterns of natural selection, recombination, or structural variation within human populations. A critical step in the analysis of such datasets is the identification of admixed or outlier individuals. These individuals need to be assigned to appropriate data subsets since incorrect identification may obscure important information or increase the occurrence of false positives. Full marker sets can be used to identify outliers, but it is more efficient to analyze a subset of maximally informative markers for each particular population in a given study. Biallelic markers are often preferred for their ease of variant detection, but random marker selection is not an optimal strategy since at least 50 markers must be used. In this issue of *BioTechniques*, Raaum et al. from the City University of New York (Bronx, NY) describe a method for selecting maximally informative biallelic markers for assigning individuals to their population of origin

by principal component analysis (PCA)-based ranking. The authors used a previously published *Alu* dataset to show that fewer *Alu* markers were required to correctly assign sub-Saharan Africans, East Asians, and Europeans to their populations of origin when using their new PCA-based method than random marker selection or a previously published PCA-based selection approach (Paschou et al. *PLoS Genet* 4:e1000114). The method also served to identify outliers from genetically and geographically intermediate populations as demonstrated by the inclusion of Indian samples with the Sub-Saharan African, East Asian, and European data. In another example, the new PCA-based method correctly identified an individual whose ancestry was admixed with adjacent geographic regions. These results indicate that the method will be valuable for identifying a minimal set of maximally informative polymorphic markers for the selection of samples from whole-genome study datasets for more in-depth analysis.

See “Efficient population assignment and outlier detection in human populations using biallelic markers chosen by principal component-based rankings” on page 449.

DNA-eating Crab

Whole-genome shotgun (WGS) sequencing of large genomes that contain high percentages of long repetitive elements—such as in higher plants—can be difficult since excessive repetitive DNA can complicate correct contig assembly from short sequence reads. Removing the highly repetitive DNA before WGS sequencing would limit wasteful sequencing of this DNA fraction of the genome and increase the amount of protein-coding sequence obtained. One approach to accomplish this is high- C_0t analysis, which is based on the renaturation kinetics of DNA. Sheared genomic DNA is heat-denatured and then slowly re-annealed. Since highly repetitive DNA elements are present at a higher initial concentration, they will reanneal and form double-stranded DNA before lower copy number DNA sequences. The double-stranded repetitive DNA can then be separated from the single-stranded low-copy-number DNA using hydroxyapatite chromatography. In this issue, E. Bogdanova and colleagues at Evrogen and the Shemiakin-Ovchinnikov Institute of Bioorganic Chemistry (Moscow, Russia) have further simplified this technique by dispensing with the physical separation of double-stranded from single-stranded DNA. Instead, they use the thermostable duplex-specific nuclease (DSN) isolated from the Kamchatka crab to digest double-stranded highly repetitive DNA after the

Initial steps in the DSN-normalization procedure for removing repetitive DNA.

appropriate annealing time. Undigested, single-stranded low-copy-number DNA is then amplified by PCR and used for WGS sequencing. The authors validated this “DSN-normalization” method on human genomic DNA due to its extensively characterized repetitive elements. In the normalized sample, repetitive elements comprised only 25% of the sequences, as compared to 40% for the non-normalized sample. The greatest reduction was in repetitive elements with the highest degree of nucleotide identity (i.e., the evolutionarily youngest elements), which indicates that their method will be well-suited to use in higher plant genomes with their highly

conserved repetitive elements. Analysis of eleven single-copy genes verified that DSN normalization did not affect the representation of unique gene sequences.

See “Normalization of genomic DNA using duplex-specific nuclease” on page 455.

CIRVing Up a Novel miRNA Assay

Quantitation of microRNAs (miRNAs) is crucial for analyzing their fundamental biological function as posttranscriptional

regulators of gene expression, as well as understanding their role in diseases such as cancer. This can be achieved through various methods, including Northern blotting, microarrays, and stem-loop reverse transcriptase PCR. In *The RNA World* special issue, L. McReynolds and colleagues at New England BioLabs (Ipswich, MA), introduce a novel and simplified method for measuring miRNA levels using the small interfering RNA (siRNA)-binding properties of the p19 protein. Derived from the Carnation Italian ringspot virus (CIRV), p19 blocks RNA interference in plants and binds dsRNA in a size-dependent but sequence-independent manner. A gel shift assay of 17-, 21-, and 25-nucleotide dsRNAs with 5' phosphates and 3' 2-nucleotide extensions demonstrated that the p19 fusion protein preferentially bound 21-nucleotide dsRNA, with an affinity in the low nanomolar range. The binding of different RNA and DNA oligonucleotides relative to the 21-nucleotide siRNA was then examined by competitive gel shifts. Neither dsDNA nor ssRNA competed for binding with 21-nucleotide siRNA, and a large excess of rRNA or tRNA did not either; this indicates that siRNA enrichment and miRNA detection would be possible in the background of cytoplasmic RNA. This was verified by the enrichment of 21-nucleotide siRNA in a 5000-fold excess of rat liver RNA using p19 fusion protein with a chitin binding domain bound to magnetic chitin beads. An assay for miRNA was then developed in which miRNA hybridized to a ³²P-labeled ssRNA probe could be detected by binding to the p19 beads. This method accurately quantitated the amount of endogenous miR-122a in rat liver total RNA when used with a standard curve generated with synthetic miR-122a. A comparison of the p19 bead assay with Northern blotting using the same labeled probe and time of film exposure showed that for the detection of a very low-abundance miRNA (miR-153) in the microfilariae of *Brugia malayi*, the p19 beads could easily detect miR-153 while the Northern blot did not.

See “Protein-mediated miRNA detection and siRNA enrichment using p19” on page xvii.

Coping with COPAS Data

The COPAS device is a flow cytometry instrument for small biological specimens 20–1500 µm in diameter such as *Caenorhab-*

ditis elegans, *Drosophila*, zebrafish, plant seeds or seedlings, or large cell clusters. Approximately 50 live organisms per second can be drawn through a fluorescence-compatible flow cell for quantification of fluorescent emissions, size determination by time-of-flight, and measurement of optical density. The instrument is commonly used in high-throughput phenotypic and genetic studies of *C. elegans*, an excellent animal model for the system due to its optical transparency and amenability to forward and reverse genetic approaches. Studies using *C. elegans* are often enabled by the use of fluorescent reporter transgenes, which may show significant variation in expression between animals. Because of this variation, COPAS sorting provides more accurate phenotypic assessment than microscopic examination while being rapid enough to screen thousands of animals each day. Although the instrument excels at rapid data collection, researchers previously lacked automated computational tools for data analysis. In this issue of *BioTechniques*, E. Morton and T. Lamitina from the University of Pennsylvania (Philadelphia, PA) describe a suite of MATLAB-based algorithms for automated statistical analysis and data comparison of COPAS data files. The suite consists of three computational tools: COPAquant, which reads single sample files, filters and extracts values and value ratios for each file, and returns a summary of the data; COPAmulti, which reads 96-well autosampling files, filters samples, graphs features across wells and plates, and performs common statistical tests to return results in a graphical format; and COPAcompare, which performs correlation analysis between replicate 96-well plates. The programs simplify the analysis of raw COPAS data by allowing users to rapidly graph their data, compare replicate plates, and identify statistical hits. While the authors demonstrated the capabilities of their new algorithms by applying them to data generated from an RNAi screen in *C. elegans*, the algorithms are customized to the standard data format output by COPAS instruments and can therefore be used for any COPAS application.

See “A suite of MATLAB-based computational tools for automated analysis of COPAS Biosort data” on page xxv.

Selected and written by Patrick C. H. Lo, Ph.D. and Kristie Nybo, Ph.D.

BioTechniques 48:433 (June 2010)
doi 10.2144/000113431

Share Your Results. Order Reprints.

Author Reprints

Make the most of your hard work by ordering reprints of your article published in *BioTechniques*. Reprints are an inexpensive and easy way to distribute your findings to students and colleagues alike.

Corporate Reprints

Leverage *BioTechniques*, the most powerful brand in the market. Reprints help support your sales effort by utilizing articles that spotlight your brand/products to educate customers at meetings and industry trade events.

More info:

www.BioTechniques.com/reprints