Genomics Data 8 (2016) 106–109

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in brief

Copy number analysis of the low-copy repeats at the primate *NPHP1* locus by array comparative genomic hybridization

Bo Yuan [a,b], Pengfei Liu [a,b], Jeffrey Rogers [a,c], James R. Lupski [a,c,d,e,*]

[a] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[b] Baylor Miraca Genetics Laboratories, Houston, TX 77030, USA
[c] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA
[d] Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA
[e] Texas Children's Hospital, Houston, TX 77030, USA

ARTICLE INFO

Article history:
Received 14 April 2016
Accepted 15 April 2016
Available online 19 April 2016

Keywords:
Array comparative genomic hybridization
NPHP1
Primate evolution
Low-copy repeats

ABSTRACT

Array comparative genomic hybridization (aCGH) has been widely used to detect copy number variants (CNVs) in both research and clinical settings. A customizable aCGH platform may greatly facilitate copy number analyses in genomic regions with higher-order complexity, such as low-copy repeats (LCRs). Here we present the aCGH analyses focusing on the 45 kb LCRs [1] at the *NPHP1* region with diverse copy numbers in humans. Also, the interspecies aCGH analysis comparing human and nonhuman primates revealed dynamic copy number transitions of the human 45 kb LCR orthologues during primate evolution and therefore shed light on the origin of complexity at this locus. The original aCGH data are available at GEO under GSE73962.

Specifications	
Organism/cell line/tissue	*Papio anubis*, *Macaca mulatta*, *Pongo abelii*, *Gorilla gorilla*, *Pan troglodytes*, *Homo sapiens*
Sex	NA
Sequencer or array type	Agilent customized aCGH with 8x60K format
Data format	Analyzed
Experimental factors	Normal
Experimental features	Copy number estimation of the human 45 kb LCRs or its nonhuman primate orthologues in the test samples using NA10851 as control
Consent	NA
Sample source location	Houston, TX USA

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73962.

* Corresponding author at: Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Room 604B, Houston, TX 77030, USA. Tel.: +1(713) 798-3723, Fax: +1(713) 798-5073.
E-mail address: jlupski@bcm.edu (J.R. Lupski).

2. Experimental design, materials and methods

2.1. Samples

Thirty-two DNA samples were tested for copy number of the 45 kb LCRs in humans or its orthologue in nonhuman primates. These included DNA samples of seven individuals from the International HapMap project (http://hapmap.ncbi.nlm.nih.gov/), three human cell lines, eight patients with homozygous *NPHP1* deletion, one baboon (*Papio Anubis*), two rhesus macaques (*Macaca mulatta*), one orangutan (*Pongo abelii*), three gorillas (*Gorilla gorilla*), and seven chimpanzees (*Pan troglodytes*). The DNA sample of NA10851, a human individual with known copy number of the 45 kb LCRs, was used as the universal control for the intra/inter-species aCGH experiments.

2.2. aCGH — design

aCGH was designed in an 8X60K format using the Agilent SureDesign website (https://earray.chem.agilent.com/suredesign/, AMADID# 032837), based on human reference genome hg19. High-density aCGH probes were used to tile the human *NPHP1* locus and flanking regions. There are two major groups of LCRs flanking the gene *NPHP1*, each group consisting of LCR pairs that are >99.6% identical [1]. Because of the high sequence similarity, we tiled aCGH

http://dx.doi.org/10.1016/j.gdata.2016.04.008

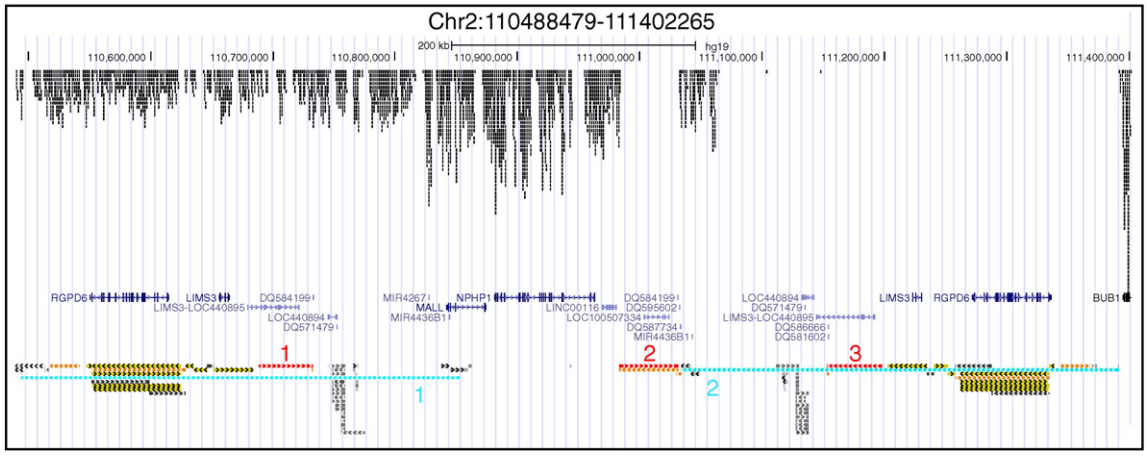

Fig. 1. aCGH design at the human *NPHP1* locus. The aCGH design is shown as a UCSC Genome Browser custom track (top) together with the tracks of UCSC genes (middle) and Segmental Dups (bottom). The tracks were aligned according to genomic coordinates in the window Chr2:110488479-111402265. As illustrated, only the LCR copies on the left side of *NPHP1* were extensively covered with aCGH probes. In the bottom Segmental Dups track, multiple LCRs including the 45 kb LCRs and 358 kb LCRs are presented for this region. We focus on the three copies of the 45 kb LCRs (highlighted in red) and two copies of the 358 kb LCRs (highlighted in blue) [1] in the haploid reference genome. The different copies of each LCR group were annotated on the Segmental Dups track with red and blue numbers, respectively.

probes from only one copy of the LCR sequences at the proximal side of *NPHP1* (Fig. 1). In this way, a clear visualization of copy number changes can be obtained by focusing on one of the two regions with high sequences similarity.

2.3. aCGH — experimental procedures

Thirty-two samples, as described above, were treated as "test samples" to compare with the universal control DNA sample of NA10851.

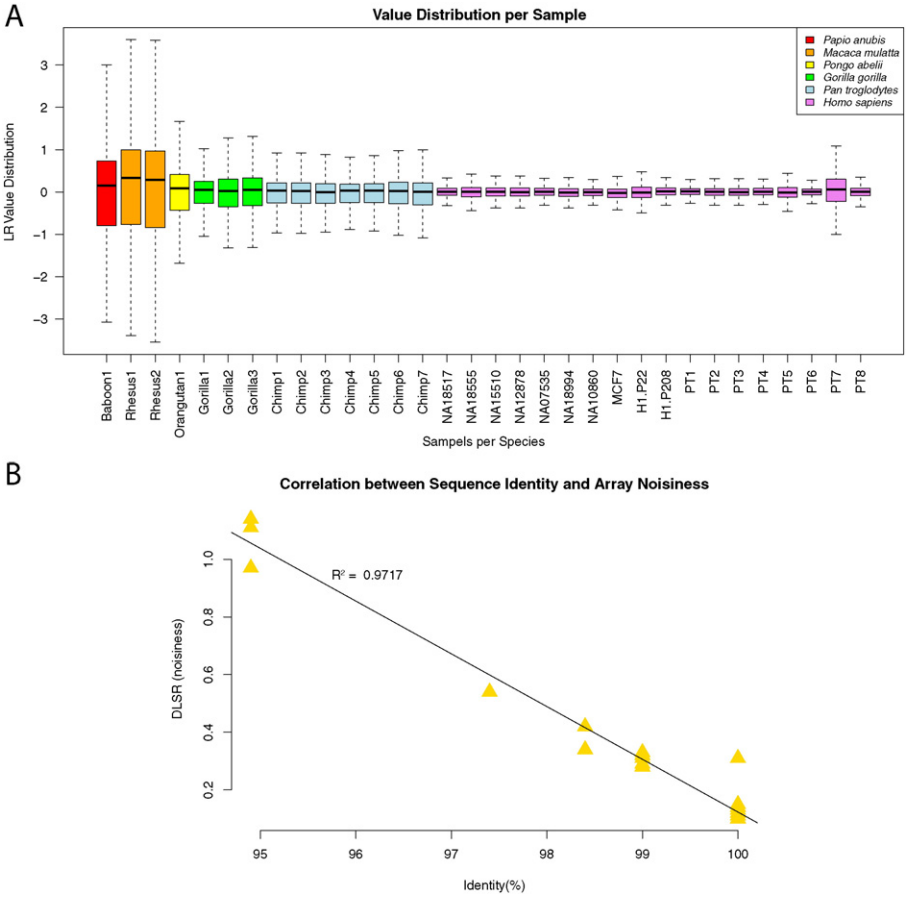

Fig. 2. Assessment of aCGH data quality. A. Value distribution of $\log_2$ ratios (LR) shown in box plots. Thirty-two samples from six species were analyzed. The wide value distribution observed in PT7 may be due to low DNA quality. The color annotations for each species are shown on the top right of the figure. B. Correlation between sequence identity and aCGH data quality. DLRS (derivative log ratio spread) is a measurement of standard deviation of the differences between adjacent points (noisiness) in log ratio data. The DLRS (Y-axis) is plotted against sequences identity (X-axis). Each gold triangle represents a sample. A strong correlation ($R^2 = 0.9717$) is observed.