

GENOME TECHNOLOGY

# Array CGH Tech Guide

**A TROUBLESHOOTING GUIDE:  
EXPERTS SHARE THEIR ADVICE ON  
PERFORMING ARRAY COMPARATIVE  
GENOMIC HYBRIDIZATION**



**M E T H O D S**

**&**

**A**



[www.nimblegen.com](http://www.nimblegen.com)

# NimbleGen CGH Microarrays & Services

## *Catch All The Breaks*



**385K    4x72K    2.1M    12x135K**

Multiple microarray formats available for analysis of a single sample, or multiple samples per slide.

- **Discover Copy Number Variants Previously Undetectable**  
Detect common and rare variants genome-wide down to ~5kb in size and fine-map breakpoints at exon-level resolution using high density 2.1M feature arrays.
- **Increase Detection of Variants in Complex Regions**  
Detect variants associated with pathogenic rearrangements using whole-genome and custom array designs with enhanced probe coverage in low-copy repeat regions of the genome (e.g. segmental duplications).
- **Choose Whole-Genome or Custom Targeted Designs**  
Catch all the breaks using whole-genome or custom targeted designs that include the most current sequence from any eukaryotic genome.

*Visit us online or call:*

[www.nimblegen.com/cgh](http://www.nimblegen.com/cgh)

**(877) NimbleGen / (608) 218-7600**



© 2008 Roche NimbleGen, Inc., Roche Applied Science  
All rights reserved. NIMBLEGEN is a trademark of Roche.

Roche NimbleGen, Inc.  
Madison, WI USA



# Table of Contents

Letter from the Editor .....	5
Index of Experts .....	5
<b>Q1:</b> How do you ensure optimal sample preparation, including DNA extraction and amplification? .....	7
<b>Q2:</b> What steps do you take to make sure you have good labeling and hybridization techniques? .....	9
<b>Q3:</b> How do you determine what type of array (BAC or oligo) to use? .....	10
<b>Q4:</b> How do you validate your results? .....	12
<b>Q5:</b> How do you ensure reproducibility? .....	14
<b>Q6:</b> What steps do you take to optimize visualization and data analysis? .....	15
List of Resources .....	18

Enter To Win an iPod Nano at [www.nanostring.com/NANOGW09.html](http://www.nanostring.com/NANOGW09.html)

Start Counting On...

Multiplexing  
High Sensitivity  
Ease of Use

...NanoString


**nanoString**  
TECHNOLOGIES

1-888-358-6266

Digital Gene  
Expression  
is What Counts

By utilizing molecular “barcodes” and single-molecule imaging, the **nCounter® Analysis System** enables the multiplexing of hundreds of gene targets in a single reaction. The nCounter System delivers high sensitivity in a fully automated, easy to use system, and is proof positive that the benefits of digital technology...are countless.

[www.nanostring.com](http://www.nanostring.com)



# Use your research resources wisely... Go Green, Go Cogenics

## The Genomics Services Company

Cogenics is setting the standard in customizing and delivering expert genomics solutions for Research, Clinical, and Manufacturing applications in the biotechnology and pharmaceutical industries.

Whether your questions are best answered by sequencing, conventional or next-generation, gene expression, genotyping, or a combination of techniques, Cogenics provides resource-effective, expertly-run solutions for your research or FDA regulated genomics projects.

Your analyses will be performed using the most appropriate platform to answer your research questions with fast delivery times and high quality data. Whether you are planning a full or pilot project, here are some of the solutions we consistently provide:

- » Sequencing solutions
- » Genetic variant assay development and validation
- » Viral and oncogene analyses
- » Drug efficacy and safety related analyses
- » SNP Discovery and Genotyping
- » Support of global multi-center clinical trials
- » Cell Bank Characterization
- » Biodistribution and Residual DNA Analyses

[www.cogenics.com/gogreen](http://www.cogenics.com/gogreen)

US: 1 877-226-4364  
UK: +44 (0) 1279-873837  
Email: [sales@cogenics.com](mailto:sales@cogenics.com)

France: +33 (0) 456-381102  
Germany: +49 (0) 8158-998518  
[www.cogenics.com](http://www.cogenics.com)

**CO:GENICS™**  
*The Genomics Services Company*

## Letter from the editor



This month, GT brings you a technical guide on array CGH. Array comparative genomic hybridization evolved from CGH, which was originally used to detect copy number gain and loss at the chromosome level. Several companies now make whole-genome microarrays for CGH that improve on this technique, offering both higher resolution and increased reproducibility.

While other detection methods have come online, including using SNP arrays to perform comparative intensity analysis, many labs have turned toward oligo arrays for CGH. BAC arrays are still used by many clinical genetics labs to

diagnose cancer and birth defects, but oligo arrays are shaping up to be a better choice for large-scale genomics research mainly because they're cheaper, easier to make, and offer higher resolution than BAC arrays.

Whatever the choice of platform, though, it's still important to nail the basics. To that end, we've compiled expert advice to address the ABCs of CGH — from optimizing DNA amplification to proper labeling, hybridization, and validation techniques. One of the main challenges to performing array CGH is data analysis, and our experts offer their suggestions on this topic, too. And for additional help, be sure to take a look at our resources section on p. 18.

— Jeanene Swanson

## Index of experts

*Genome Technology would like to thank the following contributors for taking the time to respond to the questions in this tech guide.*



**Timothy Graubert**  
WASHINGTON UNIVERSITY  
SCHOOL OF MEDICINE



**Eli Hatchwell**  
SUNY AT STONY BROOK



**Matthew Hurles**  
WELLCOME TRUST  
SANGER INSTITUTE



**Christa Martin**  
EMORY UNIVERSITY



**Steve Scherer**  
THE CENTRE FOR APPLIED  
GENOMICS, THE HOSPITAL  
FOR SICK CHILDREN, TORONTO



**Bauke Ylstra**  
VU UNIVERSITY MEDICAL  
CENTER, AMSTERDAM

## OPEN TO:

Seeing the whole genome clearly.

Resolution is everything. And when you want a highly detailed look at the genome, you need aCGH microarrays from Agilent. These microarrays feature the most robust signal-to-noise detection as well as better sensitivity and specificity than the competition. Agilent also provides a complete aCGH workflow—from sample prep to data analysis. The intuitive, user-friendly Agilent DNA Analytics software tool gives you a powerful, comprehensive view of your data in the context of the genome. If you want to take your research further, there's one thing you need to see with unparalleled clarity: The genome.

Zoom into regions of interest with

**DNA Analytics**

Agilent aCGH microarrays – raising the standard in oligo aCGH. To learn more, please visit

[www.opengenomics.com/CGH](http://www.opengenomics.com/CGH)



# Q1

## How do you ensure optimal sample preparation, including DNA extraction and amplification?

For long oligonucleotide array CGH, we have found the platforms quite forgiving with respect to sample preparation. We have not seen a significant difference in data quality using templates prepared by crude phenol:chloroform extraction vs. spin column purification. We also see equivalent results using DNA extracted from a variety of mouse tissues and from tumor vs. wild type templates. We would caution against use of whole genome amplified templates if copy number determination is the goal. In our hands, the genotype calls are highly concordant (pre- vs. post-WGA), but copy number is not always faithfully preserved during amplification.

— TIMOTHY GRAUBERT

For DNA samples we obtain in our own work (usually isolated from peripheral blood), we routinely use the Promega Wizard kit for DNA extraction. For genomic DNA samples sent to us by collaborators or customers, we check concentration, integrity, and purity by

both gel electrophoresis (to ensure that there is minimal DNA degradation) and by NanoDrop measurement, checking that the 260/280 ratio is as close to the range 1.8 – 2.0 as possible. For those gDNA samples that appear to be impure (on the basis of poor NanoDrop readings), we re-extract — our preferred method is phenol:chloroform extraction, followed by isopropanol precipitation.

There is little that can be done for samples that are heavily degraded — we try these but the results are often disappointing.

Where amounts of DNA are limiting, we favor amplification with Phi29 polymerase, using the GenomiPhi Kit from GE.

— ELI HATCHWELL

DNA is extracted from peripheral blood or tissue using a Puregene kit and the quality is checked by gel electrophoresis. If the sample is fragmented, we perform a DNA cleanup step using size exclusion columns. The quantity of DNA

obtained from the extraction is checked using a NanoDrop. Our laboratory will only proceed with microarray analysis if the DNA passes these initial quality steps. We do not amplify the samples, since we have ample genomic DNA from the peripheral blood samples that we are analyzing.

— CHRISTA MARTIN

The DNA should be isolated from the same laboratory using the same technique. Blood DNA is preferable but saliva-based samples also work well.

We maintain optimal DNA quantity and quality using NanoDrop or PicoGreen measurements for quantity and agarose gel analysis for

*continued on page 17*

---

**“Blood DNA is preferable, but saliva-based samples also work well.”**

— Steve Scherer

## Sample Prep - DNA / RNA Isolation

Experience in RNA, microRNA and DNA Isolation from:

- » FFPE Tissues as blocks, slices, or slides
- » PAXgene® and Tempus blood tubes
- » Biofluids such as saliva and plasma
- » Tissue culture
- » OCT-embedded tissues
- » Fresh Tissue
- » RNA<sup>Retain</sup>™ / RNA<sup>Later</sup>™ preserved tissues

## Expression Profiling

Leading Platforms:

- » microRNA
  - Affymetrix / Ambion DiscovArray™ Expression Service
  - ABI TaqMan® qRT-PCR assays
  - ABI TaqMan® qRT-PCR absolute quantitation
  - Agilent microarrays
- » mRNA
  - Affymetrix GeneChip®
    - Standard 2µg input
    - Proprietary 100ng Service
    - Gene 1.0 ST and Exon 1.0 ST Arrays
  - Illumina
    - Expression BeadArrays
    - DASL
  - ABI TaqMan® qRT-PCR
  - Nugen™
- » DNA
  - Agilent aCGH
  - Applied Biosystems TaqMan® SNP Genotyping assays
- » Assay Development and Validation

## Data Analysis and Bioinformatics

- » miRInform™ - Asuragen developed data delivery system for microRNA array data
  - Affymetrix/Ambion
  - Agilent
- » Biomarker Discovery
  - Feature Selection
  - Classifier Construction and Validation
- » Standard and Standard Service Premium data packages for:
  - Affymetrix GeneChips
  - Illumina Expression BeadArrays
- » Consultation / design planning
- » Statistical Analysis
- » Diagnostic Assay Characterization



## Accelerating Drug Development with Molecular Biomarkers

Asuragen provides pharmacogenomic laboratory services built on years of RNA and DNA knowledge and experience. Our capabilities can aid in accelerating drug development studies enabling our clients to focus on critical research and development activities.

Asuragen offers a wide range of unique services for quantification and data analysis of microRNA, mRNA, and DNA with sample prep services for FFPE, fresh tissues, blood, other biofluids, and cells. We are a fully licensed service provider for platforms from Affymetrix®, Illumina®, Applied Biosystems, Ambion®, and Agilent. No other service provider has the industry-leading platforms combined with so many cumulative years of experience in RNA.

We often assist our clients with experimental design and study planning so they achieve optimal information desired. Clients can use any of our service categories - sample prep, expression profiling, and data analysis, or engage us in a comprehensive project basis from isolation through data analysis. Extensive QC methods in each process provide assurance that clients will receive the best data possible from their project.

Asuragen has a proven track record in providing cGLP / CLIA services suitable for clinical testing for top pharmaceutical and biotech customers.

Please visit our website frequently for new services.



***Comprehensive miRNA, mRNA, and DNA Services***

**asuragen.com**



# Q2

## What steps do you take to make sure you have good labeling and hybridization techniques?

These steps are often performed in a core facility or contract laboratory. Quality control often includes routine UV spectroscopy, gel visualization, and assessment of yield post-labeling.

— TIMOTHY GRAUBERT

So long as the DNA quality and integrity are good, there should be no issues with DNA labeling. Our throughput is sufficiently high that our reagents tend to be fresh. On occasion, we have had difficulty with precipitates in the Cy5 dye, but we overcome this by hard spinning just before the actual hybridization (i.e., after the Cot-1 annealing).

For hybridization, it is critical to make sure that all solutions/hybridization chambers are pre-warmed. The hybridization solutions tend to be very viscous and contain nucleic acids at high concentration, making precipitation a serious concern. In the final analysis, this part of the procedure is highly dependent on the skill and experience of the individual

performing the experiment.

— ELI HATCHWELL

We follow the Agilent protocol and perform the labeling step in an ozone-free environment. After labeling, the DNA is purified using Microcon YM-30 filters and analyzed using a NanoDrop spectrophotometer to determine yield and labeling efficiency. We use opposite sex normal controls (a pool of five individuals, either male or female) for each hybridization performed. During microarray analysis, the sex chromosomes are used as our internal hybridization control; if the array shows the expected gain and loss of the sex chromosomes (gain of X and loss of Y in a female patient or loss of X and gain of Y in a male patient), then the array data can be analyzed.

— CHRISTA MARTIN

We use experienced staff and minimize the number of people that are involved in a particular protocol or experiment. We also use vendor-provided

kits, follow protocol guidelines strictly, and use liquid handling robotic instrumentation for consistency and accuracy.

— STEVE SCHERER

We always check incorporation after labeling as a last quality measure before arraying. We judge array quality by calculating the median absolute deviation of all the spots. When working with tumor samples we use a matched reference sample from the same individual when possible. This approach not only gives tighter profiles of the copy number aberrations, but profiles also devoid of copy number variations (Buffart *et al.*, 2008).

— BAUKE YLSTRA

---

**“This is highly dependent on the skill of the individual.”**

— Eli Hatchwell

# Q3

## How do you determine what type of array (BAC or oligo) to use?

Our interest has been primarily detection of copy number alterations at the highest possible resolution. For this reason, we have turned to oligo arrays. Some projects in the lab have required whole-genome views, while others have targeted specific regions of the mouse or human genomes. The flexibility of the NimbleGen custom array design is well suited to these demands.

### — TIMOTHY GRAUBERT

The choice of array is based on a set of considerations that include cost, availability, and, most importantly, available knowledge of normal variation for the platform chosen. In our case, our staple aCGH platform has been a human 19K tiling path BAC array, designed as a collaboration between my group and that of Norma Nowak at the Roswell Park Cancer Institute, and printed at RPCI. The advantage of this platform for our group is that we have data on close to 1,000 normal individuals assayed using

exactly the same platform (i.e., the 19K array). Thus, it is an easy matter for us to rapidly determine which of the CNVs we uncover in disease cohorts appear to be disease-specific and which are present in normal populations.

There is an increasing amount of data available online (especially at <http://projects.tcag.ca/variation/>) which lists structural variation discovered in normals. However, we have found that the data is patchy, with poor concordance between data elicited using different platforms and with many examples of copy number variation in supposedly normal individuals that is highly surprising (i.e., would otherwise be expected to be strongly associated with severe phenotypes).

Thus, in our opinion, it is important to possess structural variation data that has been discovered using the same platform as that used for disease studies.

The above discussion

notwithstanding, however, it is clear that the increasing resolution of aCGH afforded by emerging platforms makes these increasingly attractive. We have some experience with the NimbleGen 2.1M oligo array platform and are impressed with it (our lab was chosen as one of the beta test sites). We are also excited about trying out the new 1M feature Agilent arrays when these become available.

For region-specific analysis, where extremely high resolution is desirable, we have extensively used custom designed arrays from NimbleGen and have been pleased with the data obtained.

Clearly, another consideration in the choice of arrays is the equipment infrastructure required. For our staple BAC arrays, static hybridizations work fine (no hybridization equipment required) and a standard 5- $\mu$ m resolution Axon scanner suffices. For the new Agilent arrays, it will be mandatory to use a 2- $\mu$ m

scanner (preferably Agilent) and desirable to use a 2- $\mu$ m scanner also for the NimbleGen arrays. For Agilent, the hybridization equipment is fairly cheap while for NimbleGen, the preferred tool is a MAUI system (expensive, especially for the 12-position model — about \$50,000).

One note of caution with regard to the new generation of very high-resolution oligo arrays available from Agilent or NimbleGen: These arrays are likely to produce more data than can be interpreted rationally. Hardly any data exists on cohorts of normal individuals analyzed with these new platforms, and there are few plans to create such datasets. One company, Population Diagnostics, has as one of its stated missions to generate large sets of data for high-resolution copy number variation in normal populations of varying ethnic backgrounds.

— ELI HATCHWELL

This depends on the study, and requires us to assess the most cost efficient means of achieving the scientific objectives of the study. This not only requires that we think about the type of array, but also what array format and which supplier, because resolution and sensitivity differ between different oligo array

---

**“The trend is to move towards oligonucleotide arrays.”**

— Steve Scherer

---

suppliers. Increasingly, the higher resolution, ease of generating custom arrays, and printing reproducibility of oligo arrays is leading to the selection of these platforms for our experiments.

— MATTHEW HURLES

Our laboratory started out using BAC arrays, but quickly moved to validating oligo arrays when they became available. Oligo arrays are easier to reproduce reliably; we have noticed much more variation in the quality of BAC arrays in comparison to our current oligo arrays. In addition, with oligo arrays, it is easier to obtain a higher density of probes across the whole genome so that imbalances can be accurately sized as compared to having intervening gaps between BAC clones.

— CHRISTA MARTIN

This is the question we are most often asked. It really depends on the purpose of the study/need for resolution/available budget. No platform is perfect and each has

its strengths and weaknesses. You need to use what works for you. In a 2007 *Nature Genetics* paper we ran the same DNA sample on all available platforms and got significantly different CNV calls with each technology and CNV calling algorithm. All vendors are moving to higher resolution arrays so the data will start to stabilize, but even when using 1 million feature oligonucleotide arrays (e.g., Illumina 1M and Affymetrix 6.0) you still only see a maximum of 50% CNV call overlap.

BAC arrays are widely used in the diagnostic setting as a first screening method for exclusion of large (typically >500 Kb) cytogenetic abnormalities. Several labs have developed their own custom BAC array (spotted locally) and will therefore give preference to use it as a first tool. BAC arrays are also traditionally less noisy. These arrays are tedious to make and the trend is to move towards the easier-to-manufacture oligonucleotide arrays, but BAC arrays still have a role in clinical laboratories.

Oligo arrays can be of high probe density and/or tiling, which permits achieving high resolution compared to BAC arrays — meaning that it is able to detect smaller and more candidate CNV regions. It is the type of array preferred for research purposes, either for

*continued on page 17*

# Q4

## How do you validate your results?

This is a critical step in an aCGH experiment, especially when using oligo arrays which tend to generate somewhat noisy data. In our view, findings should be validated using orthogonal technology (e.g., PCR, SNP array, FISH). Validation should be performed on as many calls as possible, with highest priority given to the "riskiest" calls (i.e., low amplitude deviation from normal copy number, low-density probe coverage). Another critical point that has not completely permeated the literature is that detection of somatically acquired copy number alterations (or copy number neutral loss of heterozygosity) is unreliable unless matched samples from affected/unaffected tissues are directly compared.

— TIMOTHY GRAUBERT

Traditionally, we have attempted to obtain FISH validation on all the copy number variants we were interested in pursuing further. This approach, however, requires the availability of both cells

from the affected individual and a willing cytogenetics laboratory. Furthermore, FISH will not work for the validation of very small deletions or small tandem duplications (which require FISH to be more quantitative than it currently is). We have tended not to use qPCR for validation, although many people do use this approach. A 2:1 change (heterozygous deletion, for example) will be manifested by a 1-cycle difference in qPCR, while a 3:2 change (heterozygous duplication) will manifest as a ~0.5-cycle difference. Multiple replicates are required, and the CV needs to be very low for this approach to work. We favor the use of MLPA, a method we have used for some years. Historically, we have used electrophoresis-based MLPA but are currently working on Luminex bead-based MLPA, which affords greater multiplexing and does not require the use of very long oligonucleotides. Our group has developed software for the automatic

design of MLPA assays, whether for electrophoresis-based or Luminex bead-based outputs.

Homozygous deletions can clearly be validated by the use of standard PCR, which will fail to amplify the relevant sequences.

When using region-specific oligo arrays for detailed delineation of deletion/duplication/translocation breakpoints, we generally design primers that will amplify a unique junction fragment, which can then be sequenced. This provides incontrovertible validation of the structural change suspected but is limited to arrays with sufficient resolution to allow for the direct inference of junction sequences.

— ELI HATCHWELL

As there is no gold-standard reference genome or genome(s) against which we can compare results from a given experiment, we find that we generally have to generate a significant amount of validation data for each new

study. We use validation data for two subtly distinct purposes. The first is to tune the parameters in our analysis; for example, where to set CNV calling thresholds. This occurs earlier in a project. The second occurs later in a project, when we want to estimate what proportion of the CNVs identified are likely to be false positives. We think that it is important that each major survey has an unbiased estimate of their false positive rate obtained using independent validation data, so as to give users confidence and plan their experiments accordingly. Gaining an unbiased estimate of the false positive rate is not a simple procedure, not least because there is no single ideal validation technology capable of detecting the existence of all classes of CNV with a negligible false negative rate. It is important that when estimating this false positive rate in the primary CNV screen that CNVs are randomly selected for validation, rather than pre-selected on the basis of size, frequency, type, or potential biological impact. We typically use both locus-specific validation assays, such as real-time PCR using either TaqMan probes or SYBR green, and multiplexed validation assays, such as custom microarrays. For more

---

**“FISH is our preferred methodology.”**

— **Christa Martin**

---

complex variants, or for greater characterization of seemingly simple events, we use cytogenetic methods including metaphase, interphase, and fiber-FISH.

— **MATTHEW HURLES**

We validate all of our abnormal microarray results with FISH analysis, if the size of the imbalance is large enough (~100 Kb for losses and ~500 Kb for gains). If the imbalance is too small for FISH, we use qPCR, MLPA, or another array platform. FISH is our preferred methodology, since it reveals the mechanism of the imbalance (e.g., an unbalanced translocation). This information is important for recurrence risk estimates in families with a proband with a new imbalance identified by oligo array. FISH also allows performing parental testing to determine if one of the parents carries a balanced form of the rearrangement, which would not be detectable by microarray analysis since microarrays can only identify unbalanced segments of DNA.

— **CHRISTA MARTIN**

We use standard samples genotyped across labs, which permits comparison of results obtained with different platforms, array resolutions, and CNV detection algorithms. We also use replicates, or samples genotyped repeated times across time.

For validation we use non-microarray technology. Research labs will typically use qPCR or other experimental quantitative measurement (e.g., TaqMan, MLPA) by comparing the test CNV locus against a reference locus known to have two DNA copies. Clinical labs usually use FISH. We have also found using multiple programs to call CNV works well to increase discovery and help prioritize regions for validation.

How *not* to validate is by comparing to the other published CNVs (i.e., just by electronic comparison to say, the Database of Genomic Variants). You need to do some type of laboratory-based validation.

— **STEVE SCHERER**

We have used different ways to validate results, with FISH as the most common procedure. We are now in a process of moving to use Affymetrix arrays as a validation to the Agilent arrays.

— **BAUKE YLSTRA**

# Q5

## How do you ensure reproducibility?

Replicate arrays can help with this, but we have opted instead to use fewer arrays and rely on validation by other techniques (e.g., PCR/qPCR).

— **TIMOTHY GRAUBERT**

Our aCGH protocol has evolved over a period of years. Every step has been optimized. The most crucial requirement to ensure reproducibility is to stick to the protocol exactly. The first thing we teach new people in the lab who embark on aCGH experiments is to stick to the exact steps of the protocol. In our experience, most of the explanations for poor experimental data can be boiled down to variation in the way the protocol is followed. We have written down every step in great detail, so there is no need to read between the lines.

— **ELI HATCHWELL**

We can assess reproducibility both in terms of CNV calling and breakpoint estimation relatively easily through duplicate experi-

ments. Analysis of these duplicate experiments has proven invaluable in a number of studies. There are also statistical methods to allow false positive and false negative rates to be estimated from these types of data, which can be compared against empirical estimates of these parameters. Ensuring reproducibility is a different matter. We take great care to order critical reagents in large batches to minimize the batch effects. Seasonal effects, such as ozone, can be mitigated by carefully controlling the laboratory environment; for example, by installing ozone scrubbers. We monitor data quality over time and actively look for time effects. Reproducibility can also be enhanced by defining QC metrics targeted to different types of failure, and re-running failed experiments to generate a consistent final dataset. The QC metrics adopted by different companies differ substantially. We

typically use three or four QC metrics designed to capture experiments with high random noise, high systematic noise (autocorrelation), poor dose-response, and across array heterogeneity. We typically end up re-running or excluding five to 25% of experiments.

— **MATTHEW HURLES**

To minimize variation between technologists, we follow a standardized protocol developed in our laboratory that includes numerous quality control steps to check each major step of the protocol. All array processing is carried out in a controlled environment to eliminate any interfering environmental factors, such as temperature, humidity, and ozone. In addition, we are trying to automate

*continued on page 17*

---

**“Replicate arrays can help with this.”**

— **Timothy Graubert**

# Q6

## What steps do you take to optimize visualization and data analysis?

This is very much still a work in progress. Oligo array CGH data can be noisy and the datasets are very large. We and others have developed a number of algorithms to find copy number changes with high sensitivity/specificity, define boundaries with precision, and resolve complex local architecture (i.e., juxtaposition of deletions and amplifications). Currently available tools perform reasonably well, but there are still significant challenges. High on the list is the need to move from qualitative genotype calls ("normal" vs. "abnormal") to quantitative assessment of 1, 2, 3 ... copies at copy number variable regions.

— **TIMOTHY GRAUBERT**

We routinely use a 5- $\mu$ m Axon scanner and GenePix Pro. Choosing the best PMT values to use for the scan is no trivial exercise. Many people rely on the histogram to determine which values to use, but we do not favor this approach. It is important that the Cy5 and Cy3 signals are evenly matched in the features, not on the slide in general (mild

increases in Cy5 or Cy3 background can skew the histogram and suggest PMT values that do not yield balanced signals on the features). We typically choose a small region with a few representative spots and then scan at different PMTs until we find the correct values that will yield roughly equal intensities in both channels. We then apply those PMT values to the whole slide. This method works well. Historically, we relied on GenePix Pro to extract feature data and then performed manual analysis on the resulting Excel files (or used some simple macros). For the last three years, however, we have been using BlueFuse software from BlueGnome (Cambridge, UK). We favor this software for a number of reasons: the software has an algorithm which intelligently determines which features are good quality and which are not; grid alignment, feature signal extraction, fusion of data from different features with identical content, copy number calling, etc., are all automatic; and once parameters have been chosen for the

software, those parameters can be used consistently for every experiment — this ensures that data from multiple arrays can be compared to each other. In fact, we deposit all our data in a MySQL database, so that we can easily study the behavior of individual features across all arrays.

— **ELI HATCHWELL**

With each new dataset we spend quite a considerable length of time visualizing the data in different ways, to get a feel for the data and the likely sources of bias that might be minimized through normalization. Simply examining the data plotted against genomic position is a great way of visualizing the data. With noisier data, smoothing the data-points to get a sense of large-scale genomic trends has proven to be particularly useful in terms of characterizing the "wave" effect that we see in all datasets. In part, this effect results from the heterogeneous distribution of G and C nucleotides throughout the genome, and the difficulties in eradicating

subtle base composition biases in all nucleic acid-based laboratory protocols.

Visualizing the distribution of log<sub>2</sub> ratios at single probes across an entire dataset is also very useful in exploring variation in probe performance. Extracting outlier probes — for example, those with unusually high variance — and investigating reasons for these outliers is a useful first step in probe QC. This approach enables us to identify artifacts, such as autosomal probes responding to sex chromosomal content.

We use the R package extensively for most data visualization, but prefer to use C/C++ for normalization pipelines for the speed advantages. Nevertheless, R is useful for prototyping these most computationally intensive methods.

To enable these analyses we typically have to think carefully about how we store the data such that we can easily access data for all probes within a given sample as well as data for a single probe (or genomic region) across all samples. Lightweight MySQL databases have proven very useful in our work.

The model that we have adopted for data analysis typically requires a bespoke normalization pipeline to be constructed. Increasingly, this pipeline is constructed from modules that we have

---

## “Extracting outlier probes is a useful first step.”

— **Matthew Hurles**

---

used before; for example, for quantile normalization and wave correction.

Once a dataset has been finalized, sample QC is a critical step. We are typically pretty conservative. No set of QC metrics is ever perfect, and poor experiments can sometimes be best identified by examining the output of the analysis and identifying outlier samples — for example, those samples with the most/least CNV calls. Sample QC is also necessary to capture other forms of biological variation that we wish to exclude — for example, likely cell line artifacts.

There are a lot of people generating excellent software for CNV analyses, and, as well as generating our own software, we try to keep abreast of the literature. The ever-changing nature of CNV analyses requires that we take a modular approach to our analyses so as to be able to integrate new tools for individual steps in the analysis as they become available. For example, there is currently rapid growth in cross-sample CNV calling algorithms.

Once a set of CNV regions has been defined, many of the downstream analyses are very similar — for example, examining overlaps with different genomic annotations — and like most groups, we have our own in-house scripts.

For association studies we really need good statistical methods for robust association testing. If we are to adapt those methods from SNP genotyping then we need to obtain robust CNV genotypes.

— **MATTHEW HURLES**

We perform quality control analyses prior to CNV analysis. We use a powerful desktop computer for analyses (for Windows-based programs) and a Linux cluster for everything else. We organize data into databases with browser capabilities, and rank CNVs based on a prioritization list (this will depend on the project).

— **STEVE SCHERER**

For cancer research we have written many data analysis tools in the programming language R, and often integrate this with other successful array CGH bioinformatics tools developed by colleagues (van de Wiel *et al.*, 2007). For the diagnostics arrays that are analyzed by the clinical genetics department we stay with the CGH analytics, a user-friendly interface offered by Agilent Technologies.

— **BAUKE YLSTRA**



**Q1:** Continued from page 7

DNA degradation.

— **STEVE SCHERER**

We work a lot with formalin-fixed paraffin-embedded material. An overnight incubation of the isolated DNA

with NaSCN seems to be beneficial there for the final array results. We use the NanoDrop spectrum to assess if protein or phenol contaminants can be detected in the sample, and if necessary we do another cleanup using

phase-lock gels and an additional precipitation. For the FFPE material we routinely perform isothermal whole genome amplification as a DNA quality assessment (Buffart *et al.*, 2007).

— **BAUKE YLSTRA**

**Q3:** Continued from page 11

genome-wide screens or as custom array for candidate region fine-mapping (i.e., follow-up of a collection of potentially interesting CNV regions). We'll surely start to see more and more labs wanting to have a custom oligo array for screening of candidate gene regions for a particular syndromic disease or group of diseases (e.g., an array for cancer-related genes, an array for autoimmune disorders, or

an array for neurological/neuropsychiatric disorders).

For purposes of CNV association analysis, in general, oligo arrays are becoming increasingly cheaper, and many labs will be able to afford them, so they will probably slowly replace the BAC in the future. There will soon be specialized arrays with high probe coverage of common CNVs allowing CNV association testing in common diseases.

Note that the use of array

technology doesn't replace karyotyping and FISH for detection of balanced structural chromosome changes (e.g., inversions and translocations).

— **STEVE SCHERER**

We always go for the highest resolution possible, and in that respect oligo arrays outperform BACs. Since 2006, our lab has no longer produced BAC arrays (Coe *et al.*, 2007; Ylstra *et al.*, 2006).

— **BAUKE YLSTRA**

**Q5:** Continued from page 14

as much of the array processing procedure as possible. We recently introduced the use of a Little Dipper for the post-hybridization array washes. Software analysis settings and guidelines are globally set in the laboratory so that all analyses are performed using the same parameters.

— **CHRISTA MARTIN**

We ensure reproducibility by reducing error/variability and ensuring consistency between experiments. Following proto-

cols and including blind duplicate samples are key. If performing CGH, we try to use the right competitive hybridization sample.

Randomization of experiment/study design to reduce batch effects is important. For example, for family-based studies, it is ideal to have the whole family genotyped with the same batch of reagents; the same applies for case-control association studies. One 96-well plate of submitted samples would be filled with an equal number of cases and controls — half-filled with

cases, half-filled with controls. One also needs to evaluate the quality of a CNV algorithm before applying it to study samples, by, for example, randomly picking detected regions and validating them experimentally.

Ideally, results would come from one single analysis method, but no single analysis method is perfect. The more methods used, the better for discovery. A compromise is to prioritize on calls detected by at least two algorithms in order to reduce the amount of false positive calls.

— **STEVE SCHERER**

# List of resources

Our panel of experts referred to a number of publications and online tools that may be able to help you get a handle on array CGH. Whether you're a novice or a pro at the CNV game, these resources are sure to come in handy.

## PUBLICATIONS

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.** *Science*. 1992 Oct 30;258(5083):818-21.

Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nat Genet*. 1998 Oct;20(2):207-11.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet*. 1999 Sep;23(1):41-6.

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurler ME, Feuk L. **Challenges and standards in integrating surveys of structural variation.** *Nat Genet*. 2007 Jul;39(7 Suppl):S7-15.

Snijders AM, Nowak NJ, Huey B, Fridlyand J, Law S, Conroy J, Tokuyasu T, Demir K, Chiu R, Mao JH, Jain AN, Jones SJ, Balmain A, Pinkel D, Albertson DG. **Mapping segmental and sequence variations among laboratory mice using BAC array CGH.** *Genome Res*. 2005 Feb;15(2):302-11.

Snijders AM, Nowak N, Seagraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D,

Albertson DG. **Assembly of microarrays for genome-wide measurement of DNA copy number.** *Nat Genet*. 2001 Nov;29(3):263-4.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, Cremer T, Lichter P. **Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances.** *Genes Chromosomes Cancer*. 1997 Dec;20(4):399-407.

## WEB SITES

<http://cancer.ucsf.edu/array/analysis/index.php>

<http://flintbox.ca/technology.asp?Page=706>

<http://sigma.bccrc.ca/>

## DATABASES

**Center for Information Biology Gene Expression Database (CIBEX)**  
<http://cibex.nig.ac.jp/index.jsp>

**Coriell Cell Repositories NIGMS Human Genetic Cell Repository**  
<http://locus.umdj.edu/nigms/>

**Database of Chromosomal Imbalance and Phenotypes in Humans using Ensembl Resources (DECIPHER)**  
<http://www.sanger.ac.uk/PostGenomics/decipher/>

**Human Segmental Duplication Database**  
<http://projects.tcag.ca/humandup/>

**Human Structural Variation Database**  
<http://humanparalogy.gs.washington.edu/structuralvariation/>

**NCBI Single Nucleotide Polymorphism Database (dbSNP)**  
<http://www.ncbi.nlm.nih.gov/projects/SNP/>

**Segmental Duplication Database**  
<http://humanparalogy.gs.washington.edu>

# nexus copy number<sup>tm</sup>

remove the roadblocks



- Process THOUSANDS of arrays from ANY platform
- Raw data to biological insights with just two mouse clicks
- Integrate copy number, expression and miRNA data

Get your free trial at  
[www.biodiscovery.com/no-road-blocks](http://www.biodiscovery.com/no-road-blocks)



# Evolving?

Don't change jobs without us.



E-mail your updated address information to [evolving@genomeweb.com](mailto:evolving@genomeweb.com).  
 Please include the subscriber number appearing directly above your name on the address label.

Genome Technology



“i can

publish in record time.”

“Working with one assistant, I was able to go from installation of an Illumina Genome Analyzer to publication in *Nature Methods* in just three months. The system’s automation and massive throughput have had a huge impact on my research.”

Dr. Yuan Gao  
Assistant Professor  
Center for the Study of Biological Complexity &  
Department of Computer Science  
Virginia Commonwealth University

Power. Speed. Simplicity. The Illumina Genome Analyzer puts publishable results into your hands. Quickly.

~~Next-gen~~ Sequencing  
now

[www.illumina.com/sequencing?gt](http://www.illumina.com/sequencing?gt)

SEQUENCING  
GENOTYPING  
GENE EXPRESSION

