Copy-Number Variants Detection by Low-Pass Whole-Genome Sequencing

Zirui Dong,[1,2,3,7] Weiwei Xie,[3,4,7] Haixiao Chen,[3,4] Jinjin Xu,[3,4] Huilin Wang,[1,2,5] Yun Li,[3,4] Jun Wang,[3,4] Fang Chen,[3,4] Kwong Wai Choy,[1,2,6,8] and Hui Jiang[3,4,8]

[1]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China
[2]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
[3]BGI-Shenzhen, Shenzhen, China
[4]China National Genebank-Shenzhen, BGI-Shenzhen, Shenzhen, China
[5]Bao'an Maternal and Child Health Hospital, Shenzhen, China
[6]The Chinese University of Hong Kong-Baylor College of Medicine Joint Center for Medical Genetics, Hong Kong, China
[7]These two co-authors contributed equally.
[8]These two co-authors are co-corresponding authors (richardchoy@cuhk.edu.hk; jianghui@genomics.cn).

Emerging studies have demonstrated that whole-genome sequencing (WGS) is an efficient tool for copy-number variants (CNV) detection, particularly in probe-poor regions, as compared to chromosomal microarray analysis (CMA). However, the cost of testing is beyond economical for routine usage and the lengthy turn-around time is not ideal for clinical implementation. In addition, the demand for computational resources also reduces the probability of clinical integration into each laboratory. Herein, a protocol providing CNV detection from low-pass, whole-genome sequencing (0.25×) in a clinical laboratory setting is described. The cost is reduced to less than $200 USD per sample and the turn-around time is within an acceptable clinically workable time-frame (7 days).

Keywords: copy-number variants • low-pass whole-genome sequencing

How to cite this article:
Dong, Z., Xie, W., Chen, H., Xu, J., Wang, H., Li, Y., Wang, J., Chen, F., Choy, K. W., & Jiang, H. (2017). Copy-number variants detection by low-pass whole-genome sequencing. *Current Protocols in Human Genetics*, *94*, 8.17.1–8.17.16. doi: 10.1002/cphg.43

INTRODUCTION

DNA copy-number variants (CNVs) are known to be associated with the pathogenicity of a variety of human disorders due to loss or gain of the dosage-sensitive genes (Kearney, Thorland, Brown, Quintero-Rivera, & South, 2011; Tang & Amon, 2013; Jonas, Montojo, & Bearden, 2014; Dong et al., 2016). Chromosomal microarray analysis (CMA) serves as the first tier genetic analysis for individuals with developmental disability or congenital anomalies (Choy et al., 2010; Miller et al., 2010; Cao et al., 2016), and its sensitivity depends upon probe density within the targeted region. Recent studies show next-generation sequencing (NGS) as an alternative state-of-the-art technology for improved detection of chromosomal abnormalities in clinical samples (Li et al., 2014; Liang et al., 2014; Lui et al., 2015).

CNV detection with NGS data is based on two methods: (1) counting the read depth difference between the targeted region and the flanking region, or (2) identifying the

Current Protocols in Human Genetics 8.17.1–8.17.16, July 2017
Published online July 2017 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cphg.43

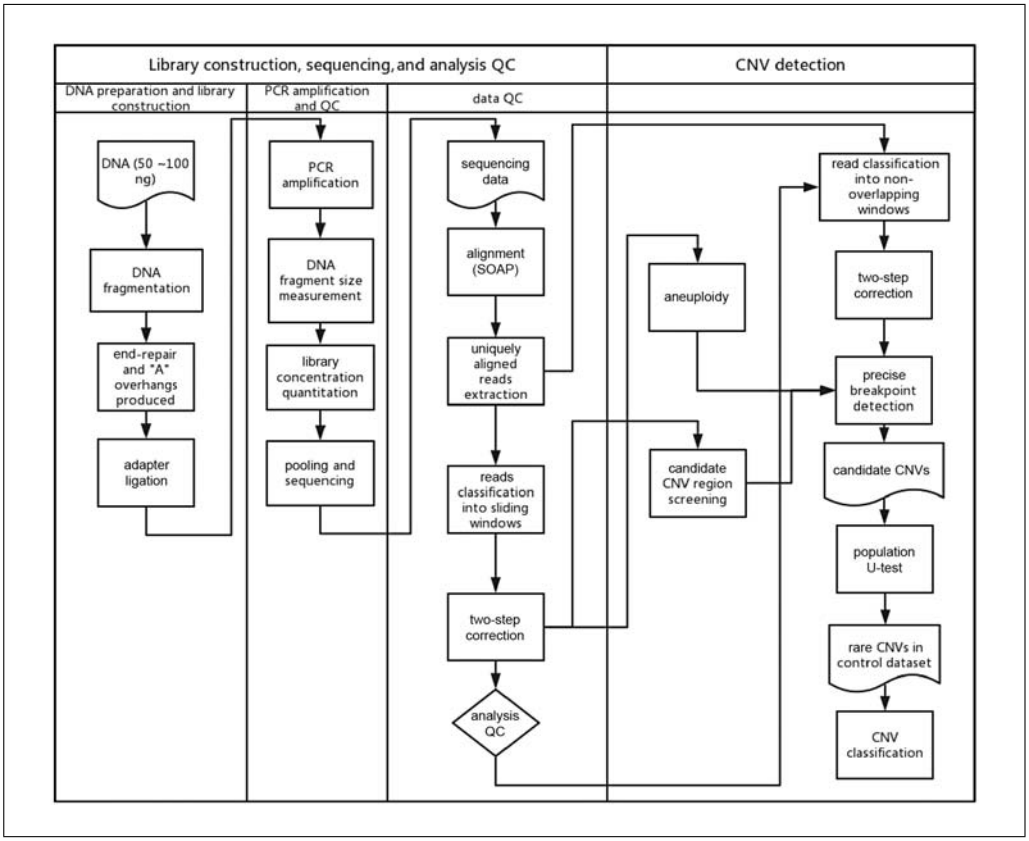

Figure 8.17.1 Workflow of CNV detection by low-pass, whole-genome sequencing. The workflow includes four protocols: (1) DNA preparation and library construction; (2) PCR amplification and library QC; (3) QC analysis, and (4) CNV detection. Detailed procedures are described in each protocol.

chimeric read pairs to support the linkage between two DNA fragments that are not contiguous in the reference genome. The first method for CNV testing was used here, as implementation of testing in a clinical setting is desired. This method only requires a small number of reads, and is therefore cost-effective and achievable within a clinical time-frame. Here, an approach (Fig. 8.17.1) for copy-number variant detection with low-pass, whole-genome, single-end sequencing, which has been validated in different clinical sample types (Dong et al., 2016), is described.

BASIC PROTOCOL 1

DNA FRAGMENTATION AND LIBRARY CONSTRUCTION

The purpose of this protocol is to create up to 96 DNA libraries suitable for single-end sequencing on a HiSeq 2000 platform (Illumina). DNA libraries for paired-end, whole-genome, sequencing require a limited size of DNA templates, which is achieved by a size-selection process, commonly gel electrophoresis; however, it is laborious and time-consuming, which is not practical in the clinical setting. To shorten the turn-around time and reduce labor costs, a protocol for rapid preparation of DNA templates for further sequencing is described here. In brief, human genomic DNA is sheared into a small insert size (i.e., 400-bp) and an adapter is ligated, while AMPure beads (Beckman Coulter) are used for size selection.

For complete testing, human genomic DNA is used as the input. A DNA sample previously used (Dong et al., 2016) was from one of the following sample types: (1) products of conception, (2) tissue from stillbirth, (3) prenatal samples, including chorionic villus, amniotic fluid, and cord blood, and (4) peripheral blood samples from paediatric cases.

In this protocol, briefly, DNA fragmentation using Covaris's Adaptive Focused Acoustics (AFA, Covaris) is used to maximize DNA recovery and restrict the DNA fragment size, and barcoded adapter ligation is performed following double-stranded DNA end repairs and the addition of adenine to the 3′ ends.

Materials

DNA sample
1% agarose gel (BioWest Agarose)
10× T4 polynucleotide kinase buffer (Illumina)
dNTP mix (Invitrogen)
T4 DNA polymerase (Illumina)
T4 polynucleotide kinase (Illumina)
Klenow fragment (Illumina)
AMPure beads (Beckman Coulter)
Elution buffer (EB) (Qiagen)
70% ethanol
10× blue buffer (Enzymatics)
dATP (Enzymatics)
Klenow exonuclease (3′-5′ exo-) (Illumina)
2× rapid ligation buffer (Illumina)
Index paired-end (PE) adapter oligo mix (Illumina) (see Table 8.17.1)
T4 DNA ligase (Illumina)
DL2000 DNA marker (Takara Biotechnology)
Lambda *Hin*dIII (Takara Biotechnology)
SYBR Safe DNA gel stain (Thermo Fisher Scientific)

Ultrasonicator (e.g., Covaris S2, Covaris)
Electrophoresis system (Thermo Fisher Scientific)
Spectrophotometer (e.g., NanoDrop 2000, Thermo Scientific)
100-µl vials (Covaris, cat. no. 520052)
1.5- and 2-ml microcentrifuge tubes (Axygen, Corning)
Vortex-5 (Haimen Kylin-Bell Lab Instruments)
Magnetic separator (Dexter Magnetic Technologies)
37°C water bath
ThermoMixer (Eppendorf)

Fragment DNA

1. Allow the ultrasonicator (e.g., Covaris S2) chiller to reach 4°C, and degas for at least 30 min.

2. During the process of chilling the ultrasonicator, prepare DNA sample:

   a. Verify integrity of DNA by 1% agarose gel electrophoresis 40 min at 150 V (Fig. 8.17.2), using the DL2000 DNA marker and lambda HindIII as size markers. Use SYBR Safe DNA gel stain for visualization of DNA.
   b. Measure purity of DNA using a spectrophotometer (e.g., NanoDrop 2000) ($OD_{260}/OD_{280} > 1.8$; $OD_{260}/OD_{230} > 1.5$).
   c. Obtain an approximate concentration using the NanoDrop 2000.
   d. Dilute 100 ng DNA to 100 µl with water and transfer DNA sample to a 100-µl Covaris vial.

3. Insert sample vial into the holder, run the Covaris with the settings given below to generate fragment sizes ranging from 200 to 350 bp:

Table 8.17.1 Genomic DNA Oligonucleotide Sequences

Name of sequence	Sequence
Barcoded adapters	
TruSeq Universal Adapter:	5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
TruSeq Adapter, Index 1	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 2	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 3	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 4	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 5	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 6	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 7	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 8	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 9	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 10	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 11	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 12	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 13	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 14	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTCCGTATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 15	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACATGTCAGAATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 16	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGTCCCGATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 18	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCCGCACATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 19	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTGAAACGATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 20	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTGGCCTTATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 21	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTTTCGGAATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 22	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTACGTAATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 23	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 25	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGATATATCTCGTATGCCGTCTTCTGCTTG
TruSeq Adapter, Index 27	5′GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTCCTTTATCTCGTATGCCGTCTTCTGCTTG
Primers	
PCR primers	5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
	5′ CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
Primer1.0	5′-AATGATACGGCGACCACCGAGATC-3′
Primer2.0	5′-CAAGCAGAAGACGGCATACGAGAT-3′
TaqMan probe	5′-CCCTACACGACGCTCTTCCGATCT-3′

Duty Cycle (%): 10
Intensity: 3
Cycles per burst: 200
Time(s): 40
Cycle: 1

Duty Cycle (or Duty Factor) is a parameter in the Covaris S2 that represents the percentage of active burst time in the acoustic treatment.

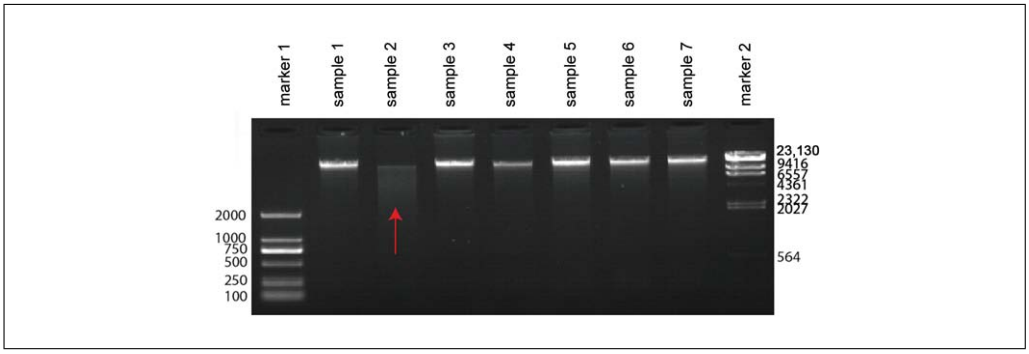

Figure 8.17.2 Verification of DNA integrity. Determine the integrity of DNAs by gel electrophoresis. Samples are shown in each column with two sets of markers (DL2000 and Lambda hindIII), size of each marker is labeled accordingly. The QC-failed sample (sample 2), which is highly degraded, is indicated by a red arrow.

4. Transfer DNA fragments into a 1.5-ml microcentrifuge tube.

Perform DNA end-repair

5. Prepare the following reaction mix:

   100 µl fragmented DNA sample (from step 4)
   0.5 µl water
   12 µl 10× T4 polynucleotide kinase buffer
   1 µl dNTP mix
   3 µl T4 DNA polymerase
   3 µl T4 polynucleotide kinase
   0.5 µl Klenow fragment
   120 µl total volume

   Mix well and spin down.

6. Incubate 30 min at 20°C using a ThermoMixer.

7. Purify with AMPure beads, as follows and elute in 34 µl EB buffer.

   a. Equilibrate AMPure XP beads to room temperature before purification. Mix the bottle of AMPure beads by gentle shaking. To generate a 120-µl end-repair reaction volume, add 120 µl of bead suspension to a 1.5-ml microcentrifuge tube.
   b. Vortex for 30 sec to achieve a homogeneous suspension and leave 5 min at room temperature (20°C).
   c. Spin down briefly and fix tube into a magnetic separator. Leave tube in the separator until solution is clear (~10 min).
   d. Remove solution by pipetting, taking care not to disturb the beads.
   e. Add 200 µl of 70% ethanol to beads, taking care not to disturb beads, leave for 30 sec at room temperature, remove ethanol by pipetting, and discard ethanol. Repeat this washing step one additional time.
   f. Leave the lid of the tube open to allow beads to dry.

      This can take longer than the recommended 5 min, but is dependent upon the volume of ethanol remaining after the previous step. The bead pellet will have a cracked appearance when it is dry.

   g. Remove tube from the magnet; add 34 µl water and mix thoroughly by pipetting. Be sure to resuspend the beads completely.
   h. Return tube to the magnet and leave for 10 min.
   i. Collect and retain the liquid (as this now contains the DNAs), and discard the beads.

Create an A-overhang

8. Prepare the following reaction mix:

   34 μl end-repaired DNA sample (from step 7)
   5 μl 10× blue buffer
   10 μl dATP
   1 μl Klenow exonuclease (3′-5′ exo-)
   50 μl total volume

   Mix well and spin down.

9. Incubate 30 min at 37°C.

10. Purify on AMPure beads (50 μl) as described above (steps 7a to 7i), and elute in 35 μl of EB buffer.

Ligate adapter

11. Prepare the following reaction mix:

    35 μl end-repaired, A-tailed DNA sample (from step 10)
    50 μl 2× rapid ligation buffer
    10 μl index PE adapter oligo mix
    5 μl T4 DNA ligase
    100 μl total volume

    Mix well and spin down.

    For each sample, randomly select one barcode adapter from the 27 listed adapters in Reagents and Solutions. However, for the samples from the same batch (12 or 24 samples), select different barcodes to distinguish each other after pooling.

12. Incubate 15 min at 20°C.

13. Purify on AMPure beads (100 μl) as described above (steps from 7a to 7i). And elute in 35 μl of EB buffer.

BASIC PROTOCOL 2

PCR AMPLIFICATION AND MEASUREMENT

After ligation with barcoded adapters, libraries are amplified by PCR with 10 to 18 cycles. It is recommended to set the number of PCR cycles as small as possible to limit amplification bias. The size of ligated DNA fragments is subsequently measured by using a 2100 Bioanalyzer DNA 1000 kit (Agilent Technologies) and library concentration is assessed by quantitative PCR (qPCR). To reduce the sequencing cost, as only 15 million reads are required for this study, in each batch, 12 or 24 libraries with different barcodes with equal molality are mixed into a pool and sequenced with 50-base, single-end sequencing on a HiSeq 2000 platform (Illumina) (Dong et al., 2016). Approximately 15 million reads are required for each sample, which is equivalent to 0.25× whole-genome coverage.

Materials

35.2 μl end-repaired, A-tailed, adapter-ligated DNA sample (see Basic Protocol 1)
Index N (N refers to barcoded adapter randomly selected from Table 8.17.1)
10× Pfx amplification buffer (Invitrogen, cat. no. 11708-013)
2.5 mM dNTP mix (Invitrogen, cat. no. R72501)
50 mM $MgSO_4$ (Invitrogen, cat. no. 11708-013)
PCR primers (see Table 8.17.1)
Platinum Pfx DNA polymerase (Invitrogen, cat. no. 11708-013)

Agencourt AMPure beads (cat. no. A29152)
EB buffer (Qiagen)
Agilent DNA 1000 kit (Agilent, cat. no. 5067-1504) containing:
  DNA dye concentrate (blue-capped vial)
  Gel matrix (red-capped vial)
  DNA marker (green-capped vial)
  DNA ladder (yellow-capped vial)
  Spin filters
  DNA chips
  Syringe
Milli-Q water
10 mM Tris·Cl, pH 8.5
0.1% and 100× Tween 20 (Sigma-Aldrich)
10× HS *Taq* buffer (Takara Biotechnology)
DMSO (Sigma-Aldrich)
Betaine (Sigma-Aldrich)
ROX (Invitrogen)
HS *Taq* (Takara Biotechnology)
Probe (10 μM)

1.5-ml microcentrifuge tubes
Thermal cycler (Thermo Fisher Scientific)
Microcentrifuge (Eppendorf)
Bioanalyzer 2100 (Agilent)
Vortex-5 (Haimen Kylin-Bell Lab Instruments)
Low-bind tubes
Magnetic separator (Dexter Magnetic Technologies)

Perform PCR

1. Prepare the following reaction mix:

   35.2 μl end-repaired, A-tailed, adapter-ligated DNA sample (from Basic Protocol 1)
   2.5 μl Index N
   5 μl 10× Pfx buffer
   2 μl dNTP
   2 μl $MgSO_4$
   2.5 μl PCR primers (see Table 8.17.1)
   0.8 μl Pfx polymerase
   50 μl total volume

   Mix well and spin down.

2. Carry out the amplification using the following cycling conditions in a thermal cycler:

1 cycle:	2 min	94°C (initial denaturation)
10 cycles:	15 sec	94°C (denaturation)
	30 sec	62°C (annealing)
	30 sec	72°C (extension)
1 cycle:	10 min	72°C (final extension)
	Indefinitely 4°C (hold)	

3. Purify on AMPure beads (40 μl) as described in Basic Protocol 1, steps 7a to 7i. And elute in 25 μl EB buffer.

Measure DNA using Bioanalyzer 2100

4. To prepare the gel/dye mix, first allow DNA dye concentrate (blue-capped vial) and gel matrix (red-capped vial), both from the DNA 1000 kit, to reach room temperature (takes 30 min). After they reach room temperature, vortex and spin down.
5. Pipet 25 μl dye concentrate into gel matrix vial, cap tube, vortex 5 sec, spin down, and transfer the gel/dye to the supplied spin filter (from DNA 1000 kit). Centrifuge spin filter 15 min at 2240 (±20%) × *g*, room temperature.
6. Position a new chip onto the priming station, and pipet 9 μl of room temperature gel/dye mix into the well marked G. Close the priming station, making sure that the syringe clip is in the lowermost position.
7. Press the plunger of the syringe until it is held by the clip and wait 60 sec.
8. Release the clip. If the plunger does not rise to ~0.6 ml within 5 sec, this is probably due to the fact that the seal has failed. If this is the case, replace the seal and repeat step 7.
9. Slowly pull the plunger back to the 1.0-ml position and open the priming station.
10. Pipet 9 μl gel/dye mix into both wells marked G.
11. Load 5 μl of marker (green-capped lid from DNA 1000 kit) into the well marked with the ladder symbol and each well that is being used to run a sample, and pipet 5 μl marker and 1 μl Milli-Q water into unused wells.
12. Load 1 μl sample into each sample well and 1 μl ladder into the ladder well.
13. Vortex on the supplied vortexer for 1 min at 2000 rpm and run the chip on the Bioanalyzer 2100.
14. Measure DNA fragment size by using the ladders and insert size of the quality control passed library should range from 250 to 350 bp.

Quantitate library concentration

The concentration of the template DNA is measured by comparing the target library to a previously sequenced library, of which a precise cluster density is known.

15. Dilute all oligonucleotide solutions to a working concentration of 10 μM with water. Use a previously sequenced library constructed with the same procedure described above and with similar distribution of fragment size, as a standard. Prepare the following three dilutions of this concentration standard in low-bind tubes: 100, 10, and 1 pM, diluting in 10 mM Tris·Cl, pH 8.5 and 0.1% Tween.
16. Dilute template DNA of unknown concentration to 10 pM in 10 mM Tris·Cl, pH 8.5 and 0.1% Tween based on their Bioanalyzer concentrations.
17. Perform qPCR assays in triplicate as follows:

    1 μl 10× HS Taq buffer
    0.1 μl $MgSO_4$
    0.1 μl 100× Tween 20
    0.5 μl DMSO
    2 μl Betaine
    0.8 μl dNTPs (2.5 mmol/liter)
    0.2 μl ROX
    0.1 μl HS *Taq*
    0.3 μl Primer1.0 (10 pM)

0.3 μl Primer2.0 (10 pM)
0.25 μl probe (10 uM)
1 μl template DNA
3.35 μl H_2O
10 μl total volume

Mix well and spin down.

18. Carry out the following cycling conditions:

1 cycle:	10 sec	95°C (denaturation)
40 cycles:	30 sec	95°C (denaturation)
	30 sec	60°C (annealing)
	45 sec	72°C (extension)

19. Adjust concentration of template DNA to 5 nM and proceed to denaturation.

20. Pool 12 or 24 samples into a library.

Using the indexed adapter tubes from a TruSeq LT Kit, 12 or 24 libraries are pooled with equal molecular amounts (pM) and subsequently sequenced with 50 single-end cycles on the HiSeq 2000 (Illumina) following the standard protocol.

BIOINFORMATIC PIPELINE

BASIC PROTOCOL 3

The analysis pipeline was described in a previous study (Dong et al., 2016). In general, the detection pipeline includes data quality control and variant detection. CNV detection is based on the read-depth difference between the targeted region and the flanking windows by combining two methods: (1) screening the candidate CNV region with adjustable sliding windows (50 kb with 5-kb increments) and (2) identifying the precise breakpoint with adjustable non-overlapping windows (5 kb) by using a module named Increment Ratio of Coverage (Dong et al., 2016). Rare CNVs implicated to be functional are identified by a population-based U-test ($P < 0.0001$). Pathogenic CNVs are further classified according to the guideline of the American College of Medical Genetics and Genomics (Kearney et al., 2011; Dong et al., 2016). Each detailed procedure described below uses a sample data as an example and the required computational resources are mentioned accordingly.

Materials

Linux-based command system
Whole-genome sequencing data (FASTQ format):
  Increment_Ratio_of_Coverage.tar.gz
  R (required module IDPmisc and SwissAir)

Download software and decompress

The latest version of the pipeline is compressed and released in a folder at *http://sourceforge.net/projects/increment-ratio-of-coverage/files/*. Run the following commands from the command line:

1. Decompress the source file:

2. `tar xzvf Increment_Ratio_of_Coverage.tar.gz` Change to the source directory (`cd Increment_Ratio_of_Coverage`):

   In this folder, the file structure is:

   i. Scripts and programs:
      a. FASTQ_fil: a script for filtering out the reads with low quality

b. SOAP2: a program for alignment (*http://soap.genomics.org.cn/soapaligner.html)*
c. rmDup: a script for removing the reads most likely generated by PCR duplication
d. Soap2Ext: a script for extracting the uniquely aligned reads and estimating the potential GC percentage of the sequenced DNA fragments (expressed as reads) with a given insert size (i.e., 500 bp)
e. Ratio_50k: a script for classification of the aligned reads into the sliding windows (50 kb with 5-kb increment) and two-step correction
f. Sample_QC: a script for analysis QC
g. Ratio_5k: a script for classification of the aligned reads into the non-overlapping windows (5 kb) and two-step correction
h. Boxplot: a script for showing the distributions of copy-ratio among different chromosomes
i. Aneuploidy: a script for aneuploidy detection
j. Increment_Ratio_of_Coverage: a program for CNV detection
k. Annotate_cnv: a script for CNV annotation

ii. Folders:
a. Config_file: a folder containing the required files for the program operations, with three sub-folders:
1. Win: this folder contains two files with the window locations' information
2. 50k: a control dataset with sliding windows; two subfolders refer to different sexes
3. 5k: a control dataset with non-overlapping windows; two subfolders refer to different sexes

b. Example: a folder containing an example for running the program and the result files

Process data and analyze quality control

3. Filter low-quality reads in FASTQ file:

```
./FASTQ_fil -fq *.fq.gz -outdir example/ -name
  example -ASII 33 -quality 5
```

This script is used for filtering out the low-quality reads, the average quality of which is 5 (default setting) with ASII encoding quality 33.

4. Align the reads to the human reference genome (hg19, GRCh37.1, hereafter called hg19):

```
./soap -a example/*.fq.gz -D hg19.fa -o
  example/*.single.soap -v 2
```

Reads were aligned to the National Center for Biotechnology Information (NCBI) human reference genome (hg19) using SOAP2 (Li et al., 2014). Six gigabytes of memory are required for each file/lane.

5. Remove reads from PCR duplication:

```
./rmDup -i *.single.soap -o *.soap.rmdup.gz -strict
  -single
```

This is used for removing the single-end reads resulting from PCR duplication. At least 3 G of memory are required for each file/lane.

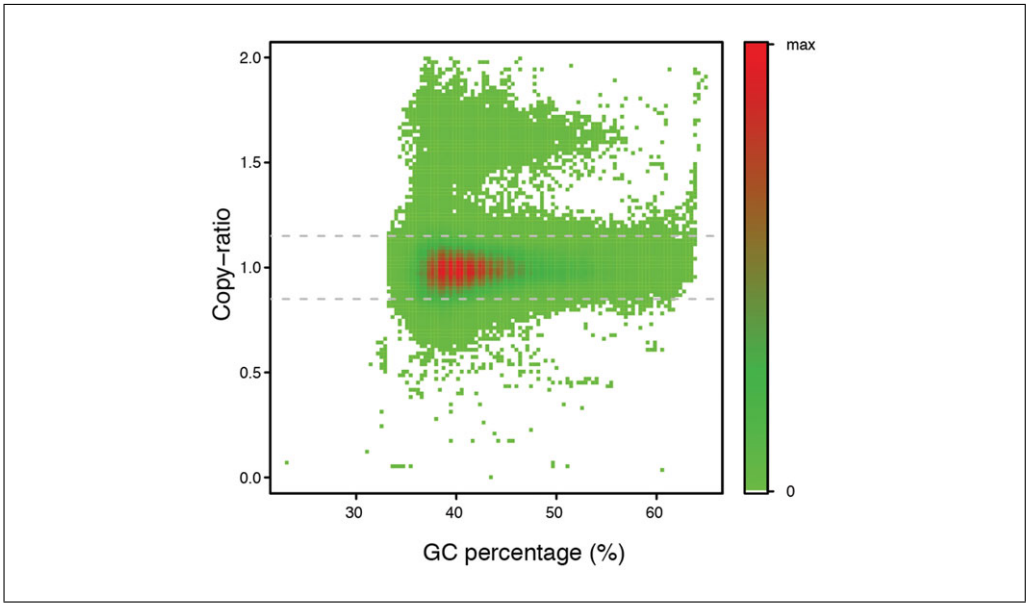

Figure 8.17.3 Quality control of the experimental samples after two-step correction. Distribution of copy-ratios (*y* axis) from sliding windows with corresponding GC percentages (*x* axis). This heat-map shows the majority of the windows having the copy-ratio as ~1 and the GC percentage as ~40%. The QC criterion is Genome-wide Standard Derivation of the windows' copy-ratios after two-step correction of <0.15. Two gray dotted lines show two cutoffs of the copy-ratios 1.15 and 0.85.

6. Extract the uniquely aligned reads and estimate the putative sex information:

```
./Soap2Ext -soap example.soap1,example.soap2,
  example.soap3 -outdir example/ -se 500 -ref
  hg19.fa -name example && sort -n -k 1 -k 2 -S 3g
  example/example.ext.record > example/example.ext
  && rm example/example.ext.record && tail -1
  example/example.gender
```

This script is used for extracting the uniquely aligned reads and defining the GC percentage of each read, which is calculated by using the reference genome sequence from the putative DNA fragment (with an insert size). It can be used to merge all of the sequencing data from a sample if more than one file/lane were generated. Three gigabytes of memory are required.

7. Classify the read with sliding windows (50 kb with 5-kb increment) and perform two-step correction (GC correction and population-based normalization) (Li et al., 2014; Dong et al., 2016); 3 G memory is required.

```
./Ratio_50k -window Config_file/win/winRatio_50k
  -ext example/example.ext.gz -outdir example/ -name
  example -gender M(or F)
```

8. Sample QC:

```
less example/example.ratio | awk '{print $5"\t"$7}'
  >example/example.ratio.cor &&./Sample_QC -i
  example/example.ratio.cor -o
  example/example.QC.svg -xlim 0.25,0.7 -xlab soapGC
  -ylab ratio -main "soapGC vs ratio" -legend
  -convert -dpi 200 × 200
```

This script is used for showing the correlation between the GC percentage and the copy-ratio of each window, as shown in Figure 8.17.3 and the file named

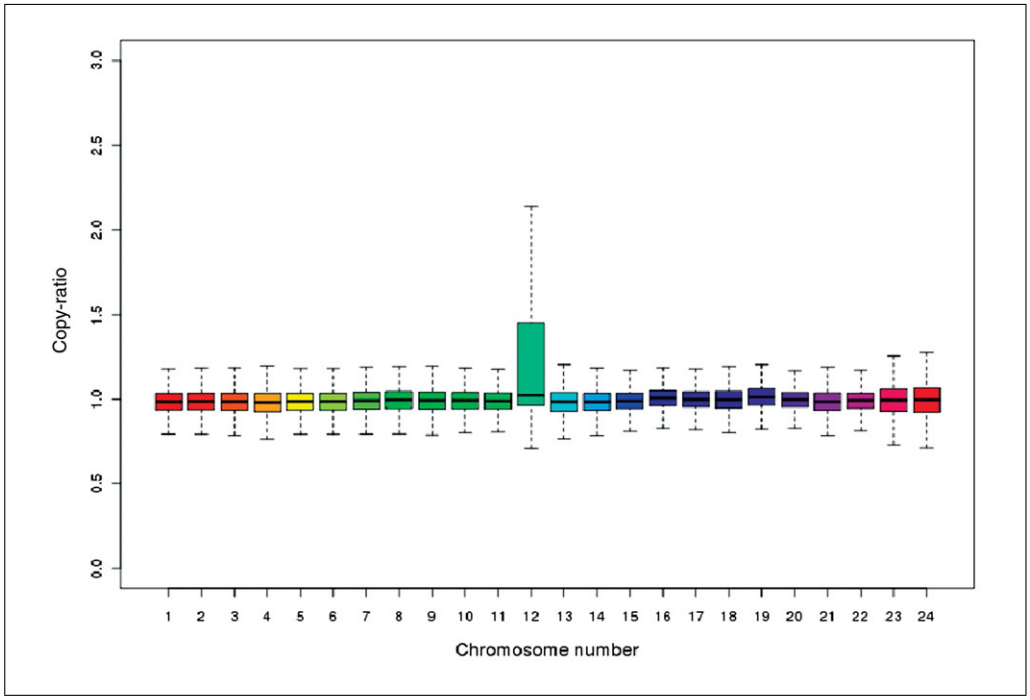

Figure 8.17.4 Copy-ratio distributions among different chromosomes in the example case. Boxplot figure shows the copy-ratio distributions (*y* axis) among different chromosomes (*x* axis) in a male subject. The copy-ratios of chromosome 12 are increased compared to the other chromosomes.

"example/example.QC.png". The authors' GC correction method is dividing the copy-ratio of a window by the correction coefficient for the windows with the same GC percentage. The main purpose is to adjust the median copy-ratio of the windows with each GC percentage into 1, reducing the false positives with over-expressed or under-expressed copy-ratios resulting from PCR amplification. As described in a previous study, the QC criterion is the genome-wide standard deviation of the windows' copy-ratios after two-step correction and should be <0.15 (Dong et al., 2016).

Detect aneuploidy

9. Classify the read with non-overlapping windows (5 kb) and perform two-step correction (GC correction and population-based normalization) (Li et al., 2014; Dong et al., 2016); 3 G memory is required.

```
./Ratio_5k -window Config_file/win/winRatio_5k -ext
  example/example.ext.gz -outdir example/ -name
  example -gender M
```

10. Detect aneuploidy:

```
./Aneuploidy -name example -i example/example.ratio
  -outdir example/ -gender F
./Boxplot -i example/example.ratio -o
  example/example.boxplot.png -g M -l 0,3
Rscript example/example.boxplot.png.R
```

These scripts are used to identify the aneuploidy, by using the median value of copy-ratios in each chromosome (germline or in mosaic fashion), as shown in Figure 8.17.4 and in file example/example.boxplot.png.

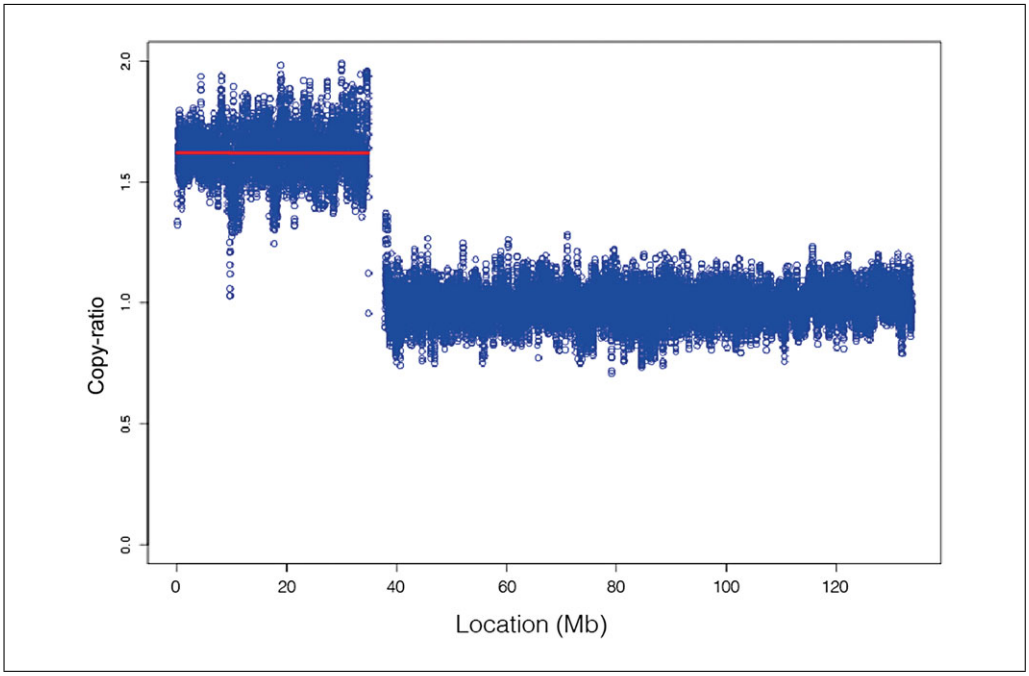

Figure 8.17.5 Micro-duplication in chromosome 12 in the example case. Distribution of copy-ratios (*y* axis) in chromosome 12 (*x* axis) in the example sample shows there is a micro-duplication of 36 Mb. The duplication is indicated with a red line.

Detect CNV and annotate

11. CNV detection:

```
./Increment_Ratio_of_Coverage -name example -outdir
  example/ -gender M -file./ -mosaic
```

This program is used to identify CNVs (Fig. 8.17.5) by combining all information generated by the scripts above. In brief, (1) report the CNV candidate region(s) with the copy-ratios from the sliding windows, (2) locally search for precise CNV breakpoints by using Increment Ratio of Coverage with the copy-ratios from the non-overlapping windows in the CNV candidate region(s), and (3) report the rare CNV(s) based on the control dataset by a population-based U-test and remove false positives by using the self-based U-test. Only 1 G memory is required.

12. Annotate CNV:

```
./Annotate_cnv -name example -outdir example/ -file./
  -clean example_clean.xls
```

The CNV classification criteria are summarized in a previous study (Dong et al., 2016) and annotation will only require 1 G of memory. Databases used for annotation are listed below.

a. RefGenes: *https://refgenes.org/*
b. GeneReviews: *https://www.ncbi.nlm.nih.gov/books/NBK1116/*
c. Online Mendelian Inheritance in Man (OMIM) (*UNIT 9.13*; Baxevanis, 2012): *https://www.omim.org*
d. Database of Genomic Variants (DGV): *http://dgv.tcag.ca/dgv/app/home*
e. ClinVar (*UNIT 8.16*; Harrison et al., 2016): *https://www.ncbi.nlm.nih.gov/clinvar/*
f. DECIPHER (*UNIT 8.14*, Corpas, Bragin, Clayton, Bevan, & Firth, 2012): *https://decipher.sanger.ac.uk/*
g. Human Gene Mutation Data (HGMD): *https://www.hgmd.cf.ac.uk/*

COMMENTARY

Background Information

In the past decade, next-generation sequencing (NGS) has been utilized as an alternative technology to chromosomal microarray analysis (CMA) that promises improved detection of chromosomal abnormalities with unprecedented resolution. Detection by low-pass, whole-genome sequencing (WGS) rather than the use of a target region–based technique is a step forward and has been proven to enhance detection particularly in probe-poor regions (Liang et al., 2014). A few retrospective studies with limited sample size have supported the performance of NGS for detecting CNVs in clinical samples (Li et al., 2014; Liang et al., 2014). In addition to germline CNV detection, identification of the mosaic CNVs and critical problems in clinical implementation, i.e., testing failure due to DNA degradation largely from fetal demise (Dong et al., 2016), have been achieved.

In this protocol, an approach for CNV detection in a clinical setting is described, the sensitivity and specificity of which have been validated methodologically and clinically (Dong et al., 2016). The resolution for CNV detection is set as 50 kb, because (1) the majority of copy-number polymorphisms are <50 kb (1000 Genomes Project Consortium, 2015) and (2) the purpose of this protocol is to develop a cost-effective approach for clinical settings (estimated to be less than $200 USD per sample), particularly in the prenatal diagnosis laboratories. Based on simulation analysis and clinical validation, this approach is robust with the described settings (i.e., 15 millions reads and 50-kb detection resolution) (Dong et al., 2016). This protocol was developed based on the HiSeq 2000 platform (Illumina), but future efforts will be directed to expand the application and implementation on other platforms, e.g., Ion Proton (Thermo Fisher Scientific).

As sequencing costs and the turn-around time decrease going forward, increased resolution of detection will be achievable at an approximately similar cost. In addition, a paired-end sequencing approach (i.e., large-insert or mate-pair library) (Dong et al., 2014; Redin et al., 2016) will be required to pinpoint breakpoints of chromosomal rearrangements and identify the exact composition of derivative chromosomes for evaluation of disrupted genes and potential positional effect predictions (Talkowski et al., 2012; Redin et al., 2016). The identification of structural variants (i.e., balanced translocations and inversions) based on the chimeric read pairs will provide a comprehensive picture of chromosomal abnormalities.

Critical Parameters

Basic Protocol 1

AFA technology is recommended for DNA fragmentation because of the restricted size of DNA fragments and higher recovery rate, compared with nebulization and sonication. For nebulization, approximately half of the DNA input is lost by vaporization and the majority of DNA fragments are of large sizes; for sonication, a relatively wide range of fragment sizes are produced.

At least 15 million reads should be generated for each sample in terms of simulation (Dong et al., 2016). Based on the experience of sequencing on the HiSeq 2000 platform (Illumina), it is possible to pool 12 or 24 samples for sequencing.

Basic Protocol 3

Only use samples passing quality control analysis for further CNV detection. CNV detection for a failed quality control sample, which may be due to DNA degradation, results in a large number of false positives because next-generation sequencing is highly sensitive to DNA fragments with high GC percentages.

The annotation databases are required to be updated but should be reformatted in terms of files stored in the Config_file folder.

Troubleshooting

After analysis, the results sometimes contain a large number of CNVs.

The same ethnic group samples tested are required, as human genomes are varied among different individuals, particularly among different ethnic populations (1000 Genomes Project Consortium, 2015). In this protocol, WGS data from ~100 Han Chinese samples (downloaded from 1000 Genomes Project, *http://www.1000genomes.org/)* were used for the control dataset in the bioinformatics pipeline.

The same sequencing platform and library construction methods are required as amplification of genomic fragments with extreme GC content is sensitive to different sequencing platforms (i.e., HiSeq and Genome Analyzer II), different library construction methods, and even different reagents (Dong et al., 2016).

Based on the method and clinical validation, a CNV database mainly based on NGS

is required for interpretation as genomes differ among individuals. In a previous study, the NGS-based approach identified 1823 benign or likely benign CNVs in the clinical groups (549 losses and 1274 gains) as defined by CMA-based databases (Dong et al., 2016).

Therefore, it is highly recommended that a control dataset be built using locally generated data. In addition, as a population-based U-test is used for detection, >30 cases for each sex are required.

Anticipated Results

If the protocols are closely followed and the control dataset generated using approximate ethnic-matched samples, a reasonable number of rare CNVs will be routinely reported with the annotated results. The sensitivity and specificity of detecting pathogenic CNVs, compared to CMA, are 100.0% (95 confident interval: 91.0% to 100.0%) and 100.0% (95 C.I.: 89.1% to 100.0%), respectively, from a previous methodological validation (Dong et al., 2016).

Time Considerations

The turn-around time from DNA preparation to CNV detection is within 1 week, estimated from a previous study (Dong et al., 2016). DNA preparation to PCR amplification requires 1 to 1.5 days. While run times for the bioinformatics analysis will vary depending on computational resources and the number of sequencing reads generated, the total run time of CNV detection is generally within 4 to 5 hr for a sample with 15 million single-end sequencing reads. The turn-around time is mainly spent on the sequencing procedure, using a HiSeq 2000 platform (Illumina). The whole turn-around time is less if a rapid sequencing platform is used (i.e., HiSeq 2500).

Acknowledgments

This study was supported by the National Natural Science Foundation of China (project 81370715), the Shenzhen Municipal Commission for Development and Reform and Key Laboratory Project in Shenzhen (CXB200903110066A and CXB201108250096A), and Health and Medical Research Fund (HMRF, Project 04152666).

Literature Cited

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi: 10.1038/nature15393.

Baxevanis, A. D. (2012). Searching online Mendelian inheritance in man (OMIM) for information on genetic loci involved in human disease. *Current Protocols in Human Genetics*, *73*, 9.13:9.13.1–9.13.10. doi: 10.1002/0471142905.hg0913s73.

Cao, Y., Li, Z., Rosenfeld, J. A., Pursley, A. N., Patel, A., Huang, J., ... Choy, K. W. (2016). Contribution of genomic copy-number variations in prenatal oral clefts: A multicenter cohort study. *Geneticae Medicae*, *18*(10), 1052–1055. doi: 10.1038/gim.2015.216.

Choy, K. W., Setlur, S. R., Lee, C., & Lau, T. K. (2010). The impact of human copy number variation on a new era of genetic testing. *BJOG*, *117*(4), 391–398. doi: 10.1111/j.1471-0528.2009.02470.x.

Corpas, M., Bragin, E., Clayton, S., Bevan, P., & Firth, H. V. (2012). Interpretation of genomic copy number variants using DECIPHER. *Current Protocols in Human Genetics*, *72*, 8.14:8.14.1–8.14.17. doi: 10.1002/0471142905.hg0814s72.

Dong, Z., Zhang, J., Hu, P., Chen, H., Xu, J., Tian, Q., ... Xu, Z. (2016). Low-pass whole-genome sequencing in clinical cytogenetics: A validated approach. *Geneticae Medicae*, *18*(9), 940–948. doi: 10.1038/gim.2015.199.

Dong, Z., Jiang, L., Yang, C., Hu, H., Wang, X., Chen, H., ... Liang, Z. (2014). A robust approach for blind detection of balanced chromosomal rearrangements with whole-genome low-coverage sequencing. *Human Mutation*, *35*(5), 625–636. doi: 10.1002/humu.22541.

Harrison, S. M., Riggs, E. R., Maglott, D. R., Lee, J. M., Azzariti, D. R., Niehaus, A., ... Rehm, H. L. (2016). Using ClinVar as a Resource to Support Variant Interpretation. *Current Protocols in Human Genetics*, *89*, 8.16.1–8.16.23. doi: 10.1002/0471142905.hg0816s89.

Jonas, R. K., Montojo, C. A., & Bearden, C. E. (2014). The 22q11.2 deletion syndrome as a window into complex neuropsychiatric disorders over the lifespan. *Biological Psychiatry*, *75*(5), 351–360. doi: 10.1016/j.biopsych.2013.07.019.

Kearney, H. M., Thorland, E. C., Brown, K. K., Quintero-Rivera, F., & South, S. T., and Working Group of the American College of Medical Genetics Laboratory Quality Assurance, C. (2011). American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Geneticae Medicae*, *13*(7), 680–685. doi: 10.1097/GIM.0b013e3182217a3a.

Li, X., Chen, S., Xie, W., Vogel, I., Choy, K. W., Chen, F., ... Zhang, X. (2014). PSCC: Sensitive and reliable population-scale copy number variation detection method based on low coverage sequencing. *PLoS One*, *9*(1), e85096. doi: 10.1371/journal.pone.0085096.

Liang, D., Peng, Y., Lv, W., Deng, L., Zhang, Y., Li, H., ... Wu, L. (2014). Copy number variation sequencing for comprehensive diagnosis of chromosome disease syndromes. *The Journal*

of Molecular Diagnostics, *16*(5), 519–526. doi: 10.1016/j.jmoldx.2014.05.002.

Lui, S., Song, L., Cram, D. S., Xiong, L., Wang, K., Wu, R., ... Yang, F. (2015). Traditional karyotyping versus copy number variation sequencing for detection of chromosomal abnormalities associated with spontaneous miscarriage. *Ultrasound in Obstetrics & Gynecology*, *46*, 472–477. doi: 10.1002/uog.14849.

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., ... Ledbetter, D. H. (2010). Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American Journal of Human Genetics*, *86*(5), 749–764. doi: 10.1016/j.ajhg.2010.04.006.

Redin, C., Brand, H., Collins, R. L., Kammin, T., Mitchell, E., Hodge, J. C., ... Talkowski, M. E. (2016). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, *49*, 36–45. doi:10.1038/ng.3720.

Talkowski, M. E., Rosenfeld, J. A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., ... Gusella, J. F. (2012). Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, *149*(3), 525–537. doi: 10.1016/j.cell.2012.03.028.

Tang, Y. C. & Amon, A. (2013). Gene copy-number alterations: A cost-benefit analysis. *Cell*, *152*(3), 394–405. doi: 10.1016/j.cell.2012.11.043.