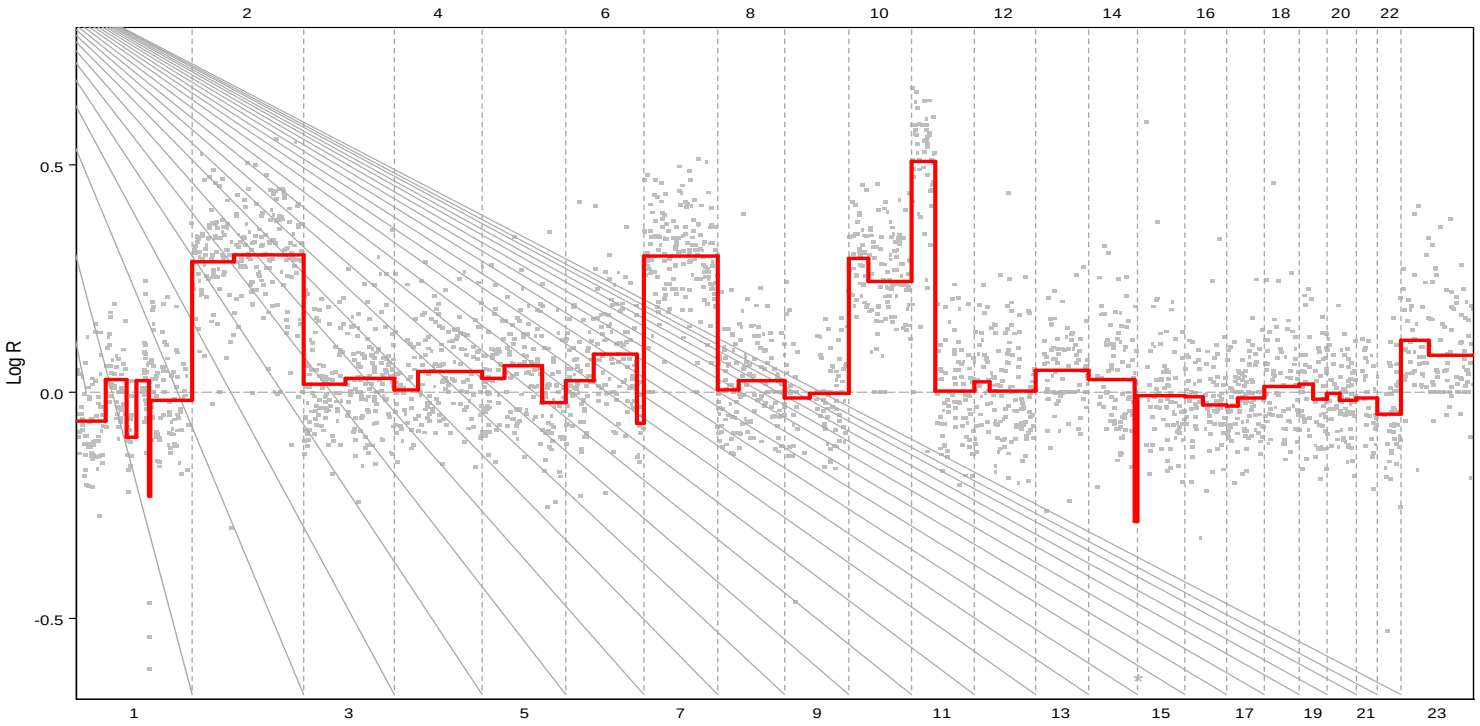
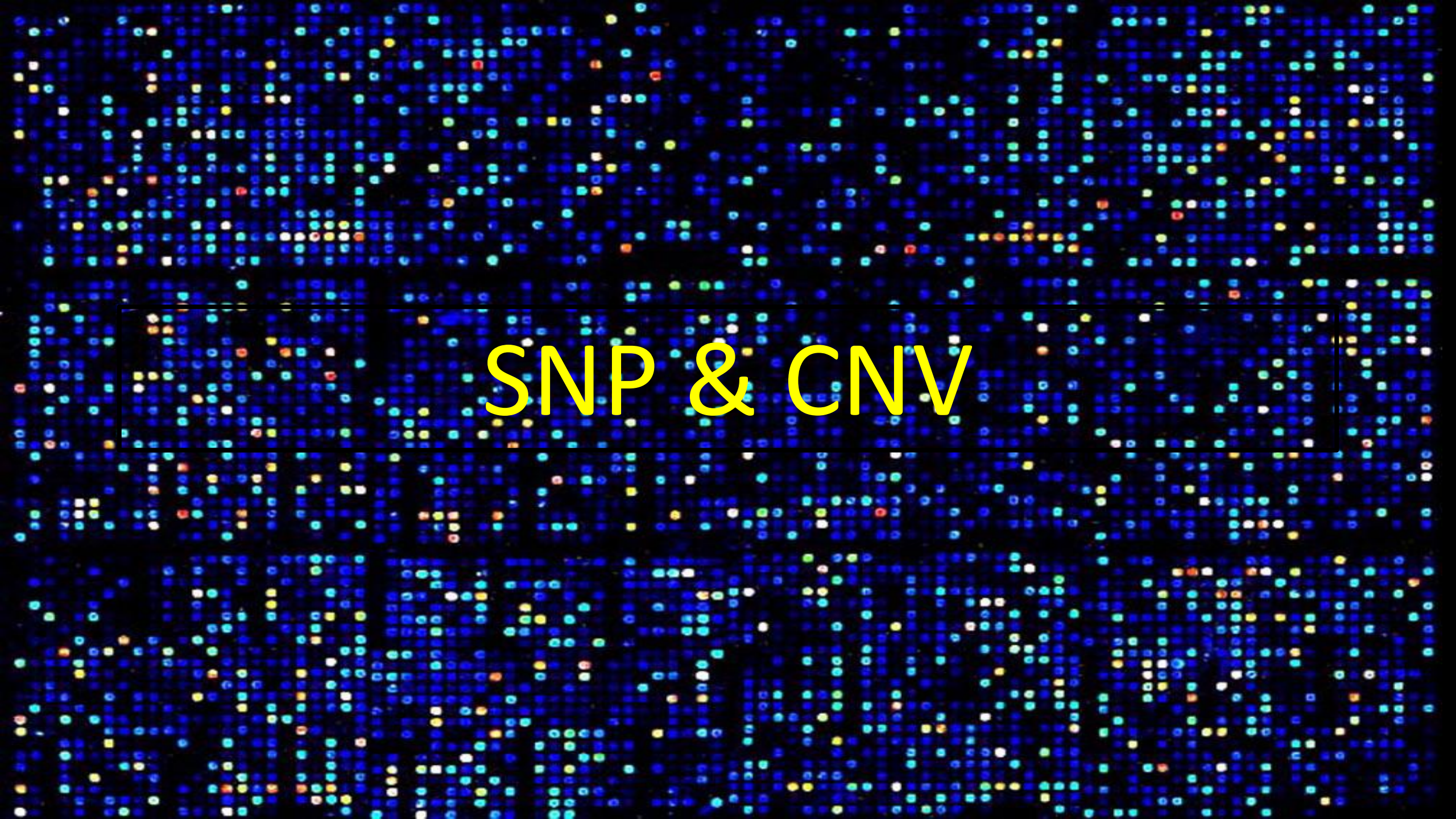
Unidad de Formación en Energías Renovables, Medioambiente y Biotecnología

TECNOLOGÍAS MICROARRAY II: COPY-NUMBER VARIATION

CURSO: **“CURSO BÁSICO DE GENÓMICA DEL CÁNCER”**





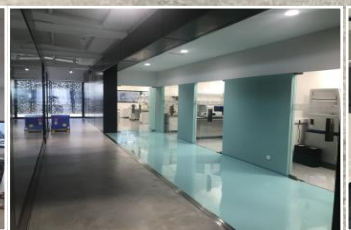
SNP & CNV

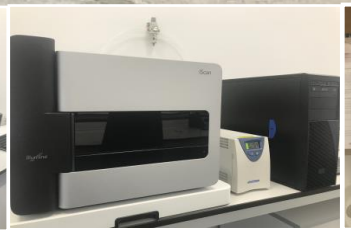
INSTITUTO DE ESTUDIOS CELULARES Y MOLECULARES

PELAYO G. DE LENA RODRÍGUEZ

pelayogonzalez@e-icm.net

COMPUTATIONAL BIOLOGIST



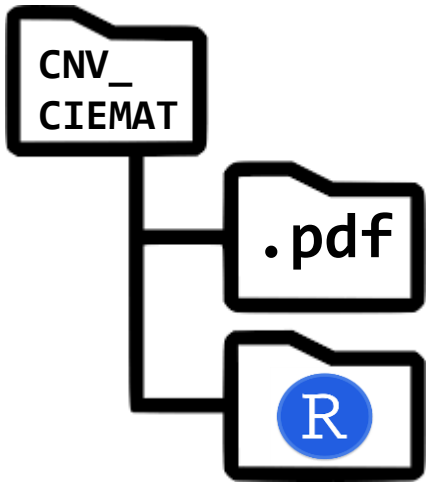


Donald Knuth
Professor emeritus of Computer Science at Stanford University
Turing Award winner
"father of the analysis of algorithms."

"I can't be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It's at that level."

Outline

- PART 1: intro --25’
- PART 2: detecting CNV --20’
- PART 3: hands on --60’



PART 1: intro to CNV

Some Terms

- **Copy number variation (CNV):** A genomic-inherited segment of at least 50 bp that differs in copy number based on the comparison of two or more genomes.
- **Copy number alteration/aberration** = somatic copy number = CNA = SCNA
- **Unbalanced rearrangements:** Genomic variants that involve loss (deletion) or gain (duplication) of segments of the genome.
- **CMA** which is also known as cytogenomic microarray analysis encompasses two different laboratory techniques, comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) arrays
- **Copy number stable(CNS):** Pertaining to regions of the genome without any detected copy number variation in healthy individuals.
- **Haploinsufficiency** :Reduction in the amount of gene product owing to functional loss of an allele that leads to an abnormal or a disease state.
- **Copy number variable regions (CNVRs):** Regions containing at least two copy number variations that overlap and that may have different breakpoints.
- **SNP-based arrays:** Single-nucleotide polymorphism (SNP)-based microarrays that contain SNP probes to genotype human DNA at the single-base level. However, through dosage signals in adjacent regions, they can be used to recognize copy number variations.

What is Copy Number Variation (CNV)?

- inherited deviations from a diploid state within a particular chromosome segment.
- >1kb

Nucleotide

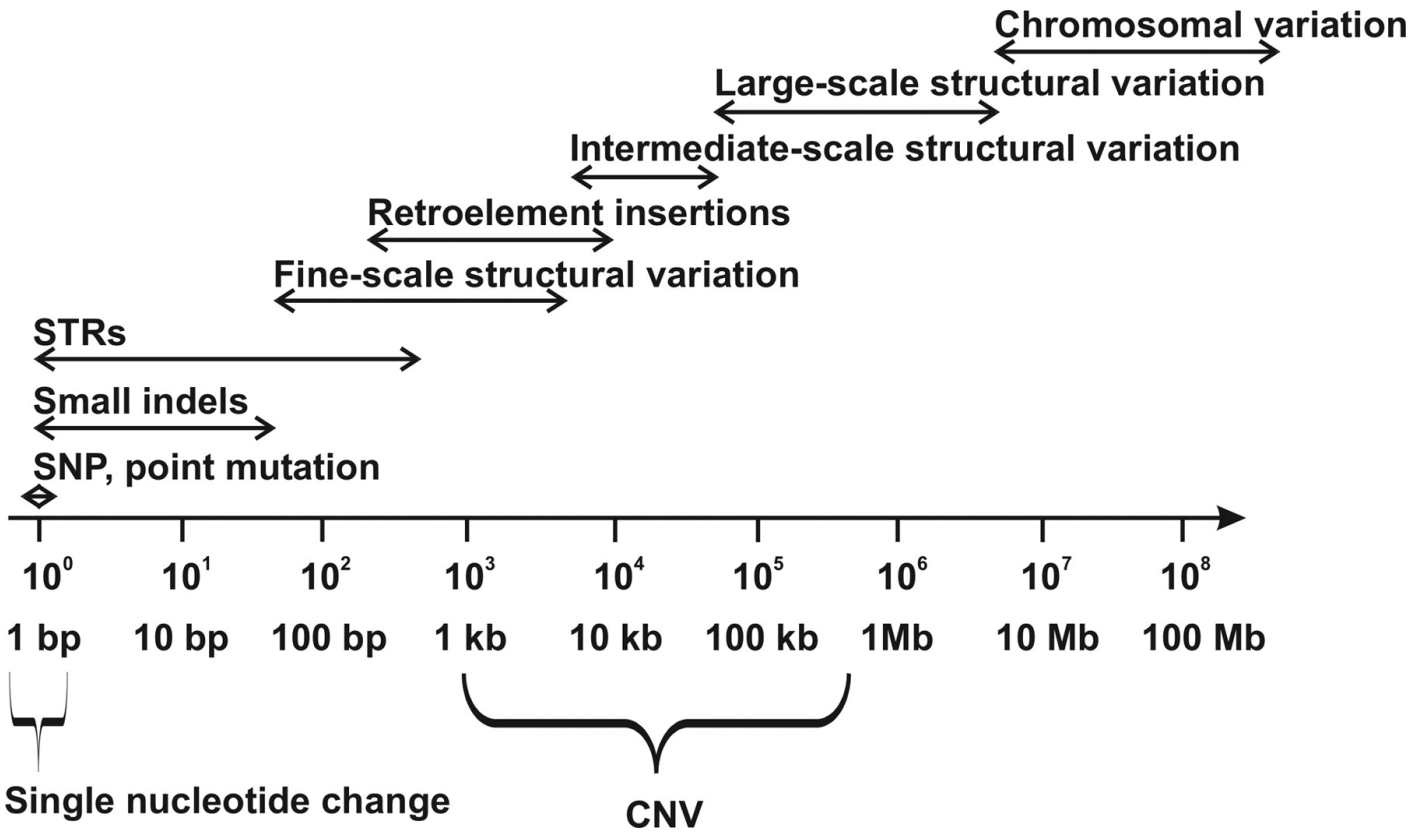
- **Single base-pair changes**
  - Point mutations (1 per 800 bp)
- **Small insertions/deletions**
  - Frameshift, microsatellite, minisatellite
- **Mobile elements**
  - Retroelement insertions (300bp -10 kb)
- **Large-scale genomic copy number variation** (>10 kb)
  - Large-scale Deletions
  - Segmental Duplications
- **Local Rearangements**
- **Chromosomal variation**
  - Translocation, inversion, fusion

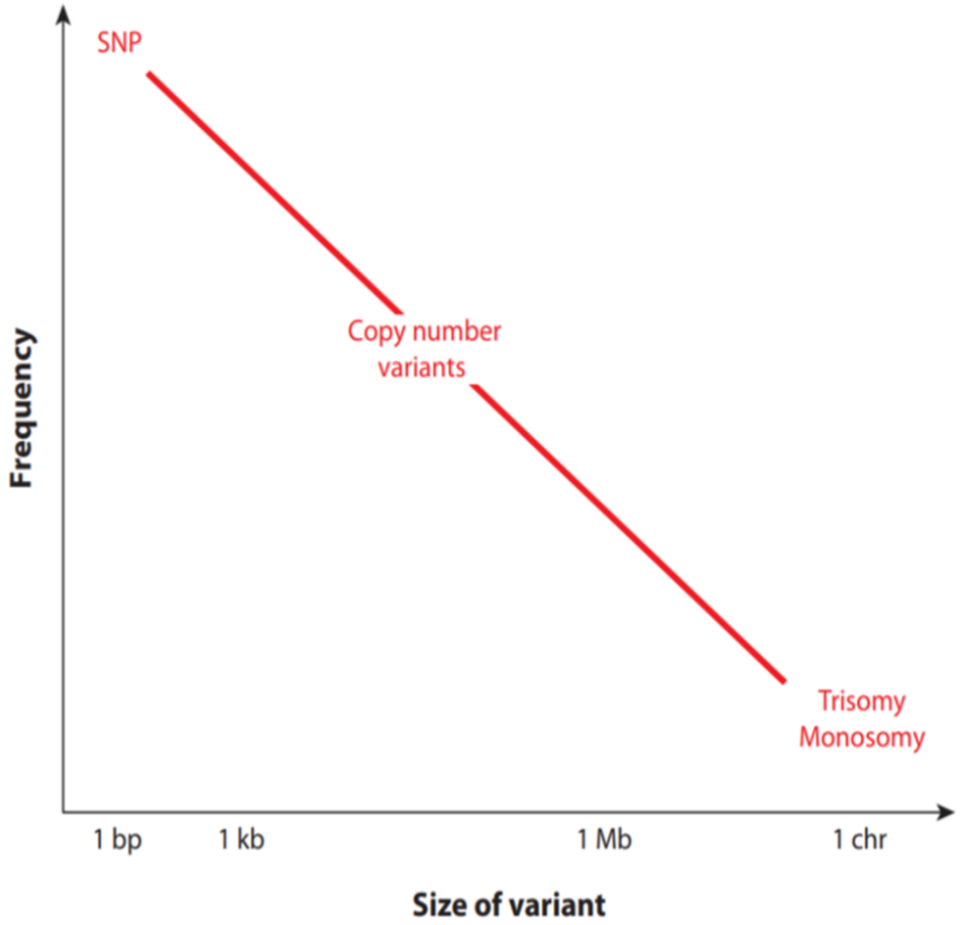
Cytogenetics

Copy Number Variation

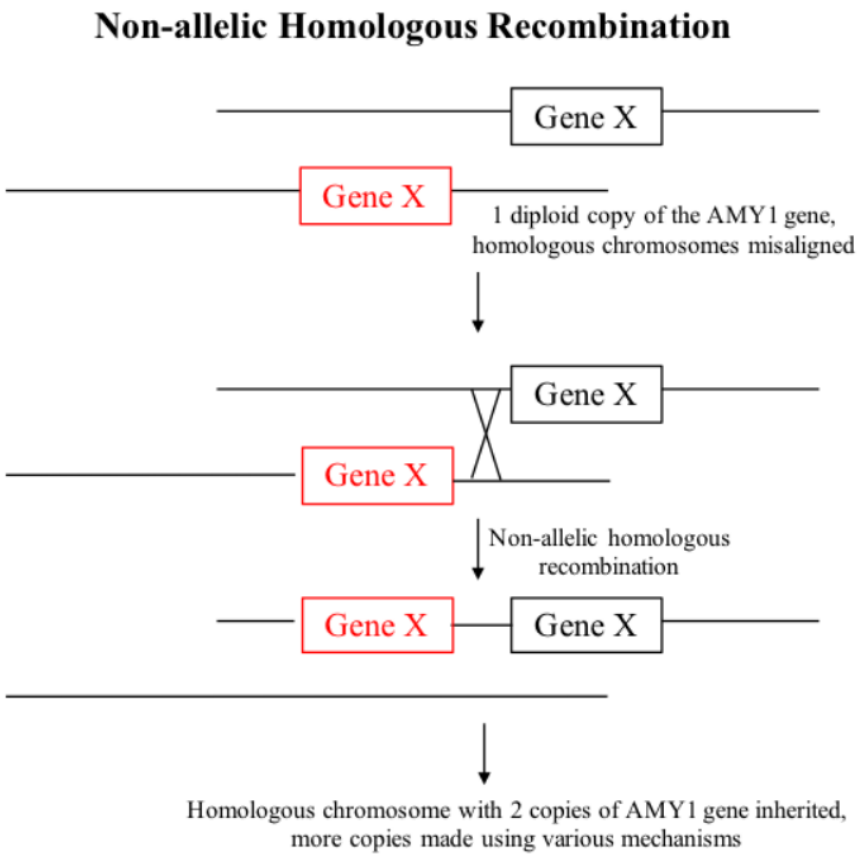
Structural Variants (SV)

What is Copy Number Variation (CNV)?





Where do CNVs cames from?



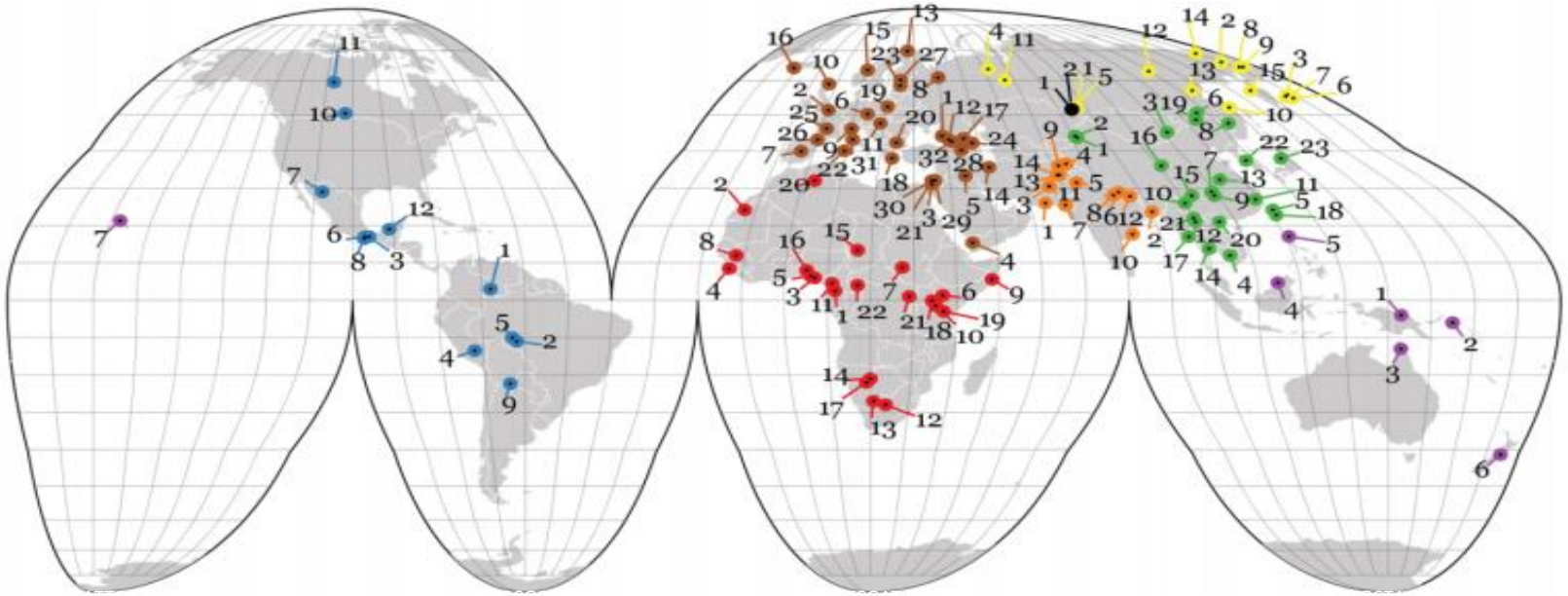
Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. Nat. Genet. 43, 838–846 (2011).

Where is Copy Number Variation (CNV)?

panel of 236 human genomes representing 125 diverse human populations

Global diversity, population stratification, and selection of human copy number variation

Peter H. Sudmant,[1] Swapan Mallick,[2,3] Bradley J. Nelson,[1] Fereydoun



1 - Neanderthal
2 - Denisova
1 - Piapoco
2 - Surui
3 - Mixe
4 - Quechua
5 - Karitiana
6 - Mixtec
7 - Pima
8 - Zapotec
9 - Chane
10 - Cree
11 - Chipewyan
12 - Maya
1 - Xibo
2 - Uygur
3 - Mongola
4 - Cambodian
5 - Atayal
6 - Oroqen
7 - Tujia
8 - Hezhen
9 - Miao
10 - Naxi
11 - She
12 - Dai
13 - Han
14 - Thai
15 - Yi
16 - Tu
17 - Burmese
18 - Ami
19 - Daur
20 - KinhKVH
21 - Lahu
22 - Korean
23 - Japanese
1 - Papuan
2 - Bougainville
3 - AustraliaECCAC
4 - Dusun
5 - Igorot
6 - Maori
7 - Hawaiian
1 - Adygei
2 - EnglandGBR
3 - Bedouin
4 - YemeniteJew
5 - IraqiJew
6 - Czechoslovakian
7 - SpainIBS
8 - Russian
9 - Bergamo
10 - Orcadian
11 - Hungarian
12 - Abkhasian
13 - Saami
14 - Iranian
15 - Norwegian
16 - Icelander
17 - Chechen
18 - Greek
19 - Polish
20 - Bulgarian
21 - Palestinian
22 - Sardinian
23 - Estonian
24 - Lezgin
25 - French
26 - Basque
27 - FinlandFIN
28 - Armenian
29 - Jordanian
30 - Druze
31 - Tuscan
32 - Georgian
1 - Kongo
2 - Saharawi
3 - Igbo
4 - MendeMSL
5 - EsanESN
6 - LuhyaLWK
7 - Dinka
8 - Mandenka
9 - Somali
10 - MasaiMKK
11 - Lemande
12 - BantuSETswana
13 - KhomaniSan
14 - San
15 - GambiaGWD
16 - Yoruba
17 - BantuSEHerero
18 - Luo
19 - BantuKenya
20 - Mozabite
21 - Mbuti
22 - Biaka
1 - Makrani
2 - BengaliBEB
3 - Balochi
4 - Burusho
5 - PunjabiPJL
6 - Tibetan
7 - Sindhi
8 - Kusunda
9 - Kalash
10 - Madiga
11 - Brahui
12 - Sherpa
13 - Hazara
14 - Pathan
1 - Tubalar
2 - EskimoNaukan
3 - AleutBeringIsland
4 - MansiSosvaRiver
5 - Altaian
6 - TlingitPreobrazhenskoye
7 - TlingitNikolskoye
8 - EskimoChaplino
9 - EskimoSireniki
10 - Ulchi
11 - MansiKondaRiver
12 - Yakut
13 - Even
14 - Chukchi
15 - Itelman

Condensed history

Gross copy number (CN) alterations were initially detected with karyotyping in the early days of cytogenetics.
(Pepler et al., 1968; Dowjat and Wlodarska, 1981; Nistér et al., 1987)

Today, the most widely used techniques can be classified as
- amplification-based (*.PCR),
- hybridization-based (FISH, comparative genome hybridization, and SNP arrays)
- sequencing-based.

CNV:Intensive Area of Research

RESEARCH ARTICLE

Intronic CNVs and gene expression variation in human populations

Maria Rigau[1], David Juan[2], Alfonso Valencia[1,3]*, Daniel Rico[4]*

1 Barcelona Supercomputing Center (BSC), Barcelona, Spain, 2 Institut de Biologia Evolutiva, Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Spain, 3 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, 4 Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom

New Results

Whole-genome sequencing analysis of genomic copy number variation (CNV) using low-coverage and paired-end strategies is highly efficient and outperforms array based CNV analysis

Bo Zhou, Steve S Ho, Xianglong Zhang, Reenal Pattni, Rajini R Haraksingh, Alexander E Urban

doi: https://doi.org/10.1101/192310

This article is a preprint and has not been peer-reviewed [what does this mean?].

Gao et al. BMC Genomics (2017) 18:265
DOI 10.1186/s12864-017-3636-3

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-00556-x OPEN

CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits

Aurélien Macé et al.#

New quality measure for SNP array based CNV detection

A. Macé[1,2,3,*], M.A. Tuke[4], J.S. Beckmann[3], L. Lin[5], S. Jacquemont[6], M.N. Weedon[4], A. Reymond[7] and Z. Kutalik[1,3,*]

[1]Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland, [2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [4]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK, [5]Division of Cardiology, Geneva University Hospital, Geneva, Switzerland, [6]Service de Génétique
r Vaudois, Lausanne, Switzerland and [7]Center for Integrative Genomics,
erland

BMC Genomics

ssed.

ccepted on July 3, 2016

RESEARCH ARTICLE Open Ac

Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans

Rajini R. Haraksingh[1,2], Alexej Abyzov[3] and Alexander Eckehart Urban[1,2,4*]

RESEARCH ARTICLE Open Access

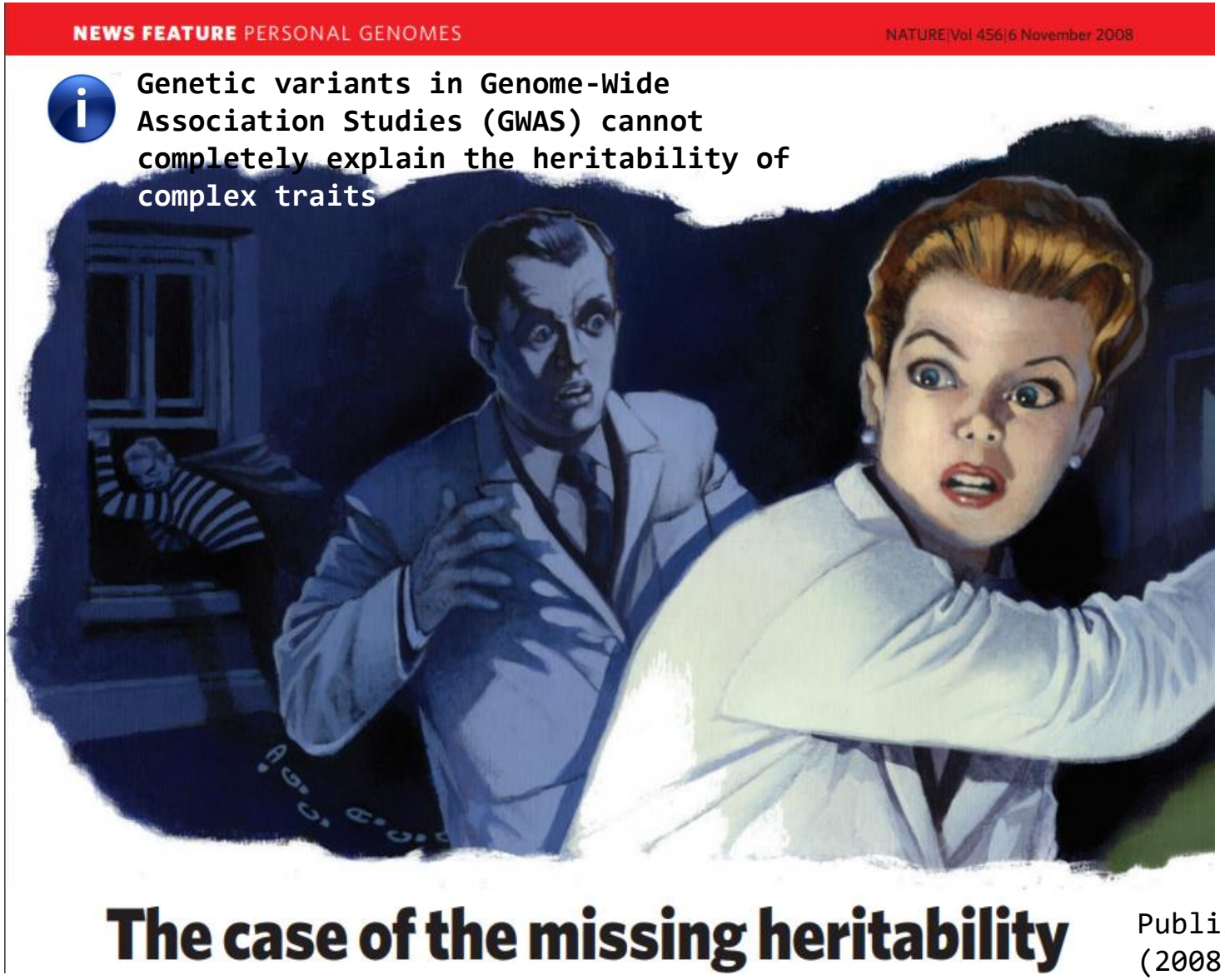
CNV discovery for milk composition traits in dairy cattle using whole genome resequencing

Yahui Gao[1], Jianping Jiang[1], Shaohua Yang[1], Yali Hou[2], George E Liu[3], Shengli Zhang[1], Qin Zhang[1] and Dongxiao Sun[1*]

ARTICLE Open Access

hensive performance comparison
resolution array platforms for
-wide Copy Number Variation
nalysis in humans

[1,2], Alexej Abyzov[3] and Alexander Eckehart Urban[1,2,4*]



Personal genomes: The case of the missing heritability

- SNPs could explain about 10-15% genotype-phenotype.
- Epistasis?
- **CNV?**
- lncRNAs?
- miRNAs?
- Diet?
- Gut?
- HLA?
- Chromotrypsis

Published online 5 November 2008 | Nature 456, 18-21 (2008) | doi:10.1038/456018a

Examples

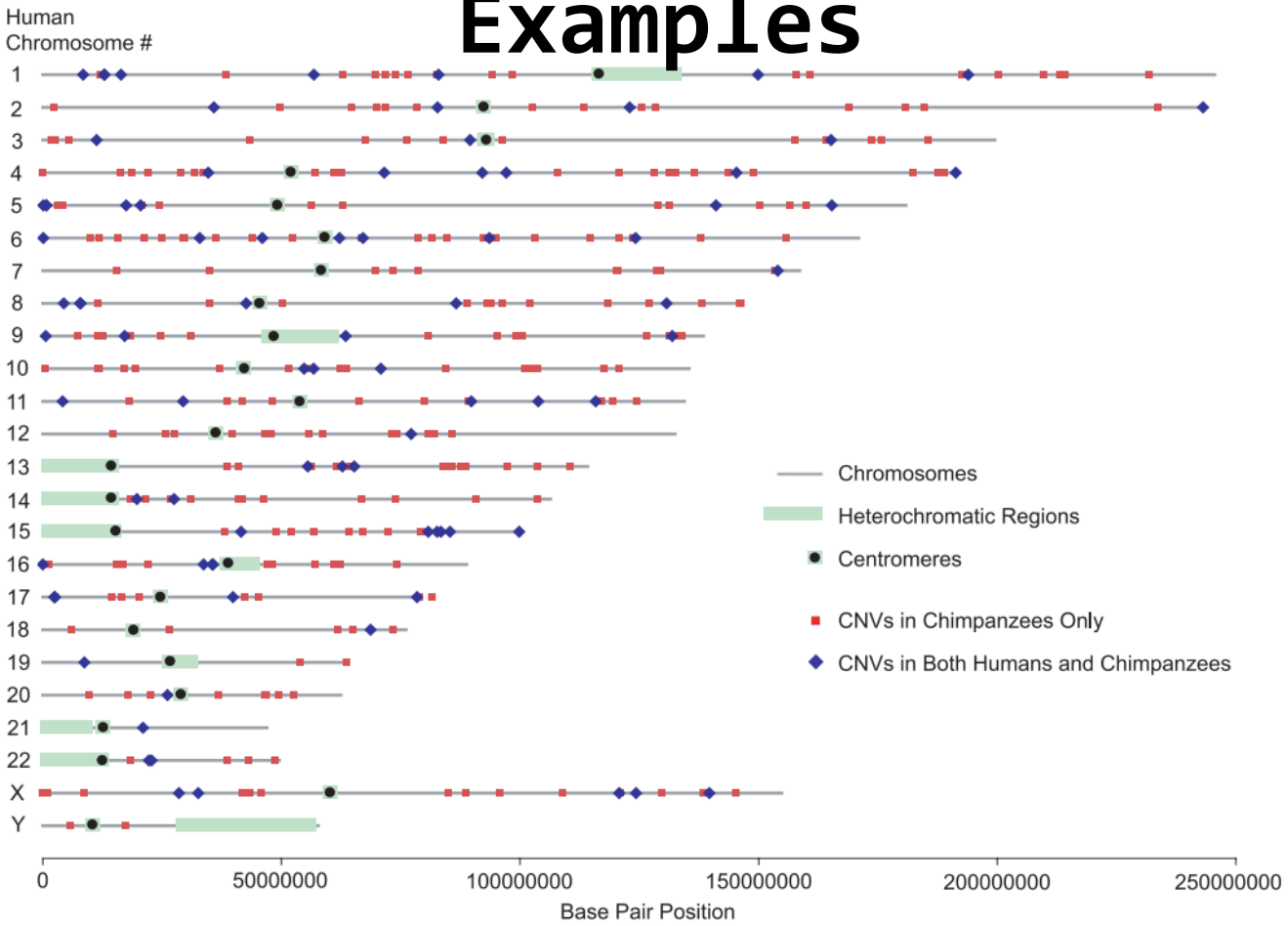
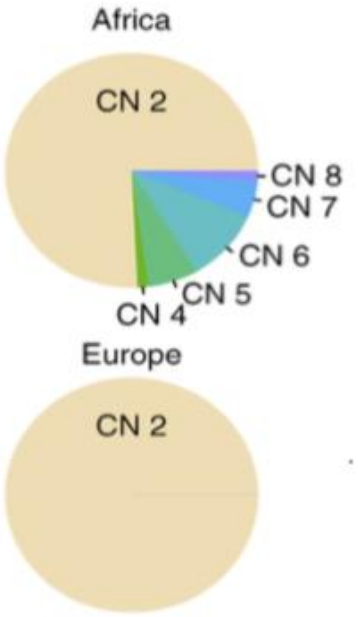


Fig. 1. Map of CNVs shared between humans and chimpanzees and found in chimpanzees only. The genomic positions of chimpanzee CNVs that do not overlap with any currently known human CNVs are depicted as red squares, and CNVs observed at the same loci in both chimpanzees and humans are depicted as blue diamonds.

Hotspots for copy number variation in chimpanzees and humans.
George H. Perry et al PNAS May 23, 2006. 103 (21) 8006-8011; https://doi.org/10.1073/pnas.0602318103

Examples

HPR gene copy number frequency

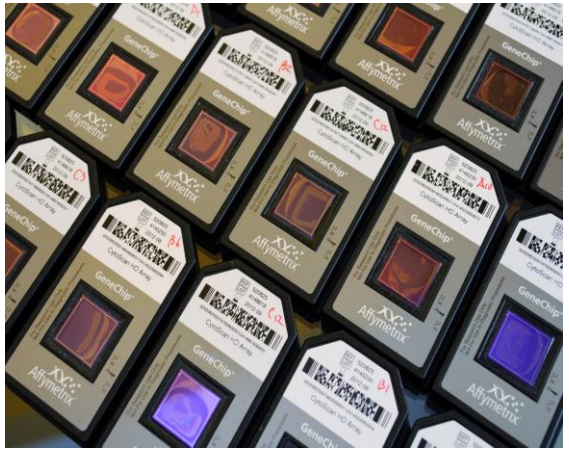


HPR encodes a haptoglobin-related protein that is used in **defense against trypanosomes.**

PART 2: detecting CNV --15’

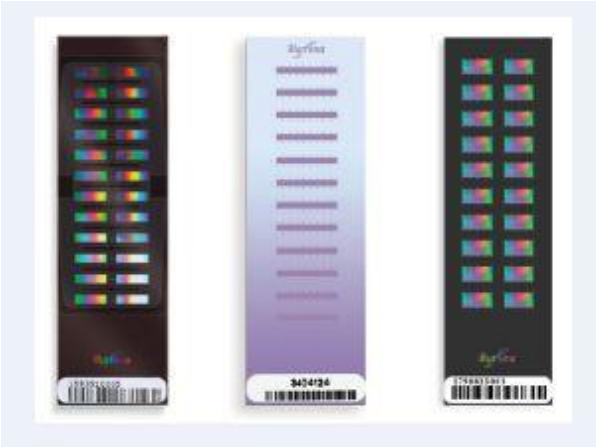
Microarray-based technologies for detecting CNV

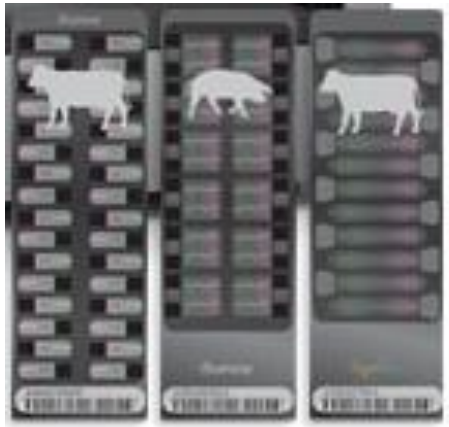
- Illumina
- Affymetrix
- Roche-NimbleGen
- Agilent



Major Market Competitors:

Illumina, Inc., Thermo Fisher Scientific Inc., Agilent Technologies, Bioline, Qiagen, Danaher, F. Hoffmann-La Roche Ltd., General Electric Company, Sequenom, FLUIDIGM, BiogeniQ Inc., Envigo, Helix, LGC Limited, Genotypic Technology Pvt Ltd, Eurofins Genomics, Biofortuna Ltd, Agena Bioscience, Inc., BioTechniques, Precision Biomarker Resources, Inc.,
TrimGen Corporation, 23andMe, Inc., Xcelris Genomics, Integrated DNA Technologies, Inc. and BioTek Instruments, Inc. among others



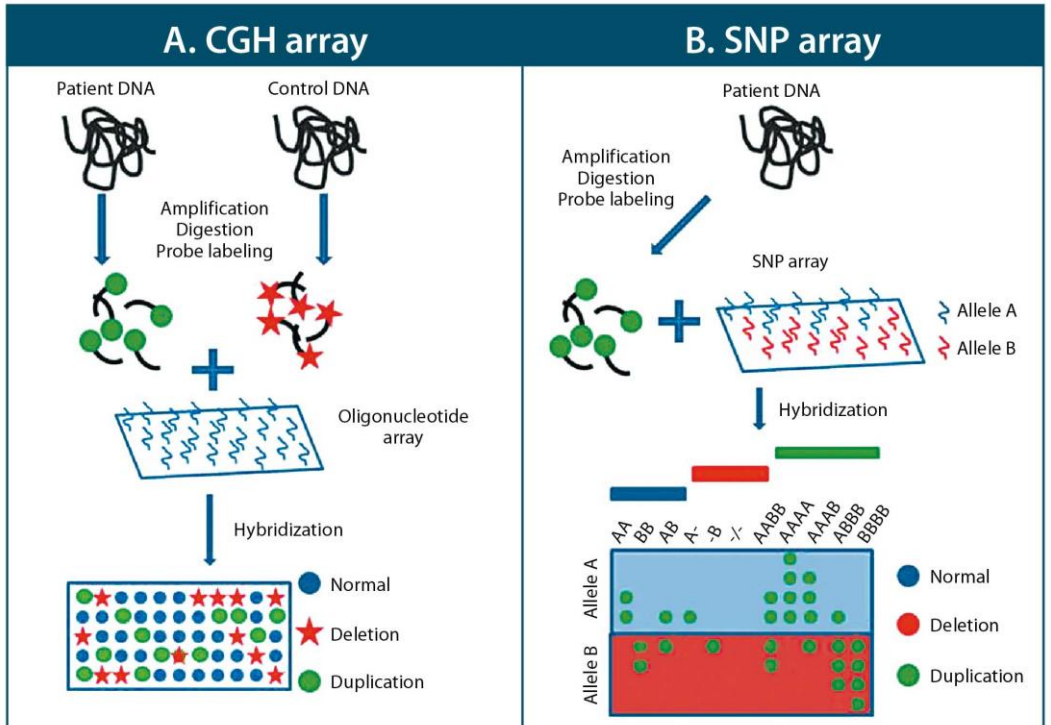


Haraksingh et al. BMC Genomics (2017) 18:321

Table 1 Features and experimental details of CNV detection platforms used in this study

Array	Total Features	Probe Spacing	Data Source	Platform-Specific Analysis Software	Number of Autosomal CNVs Called (replicate 1, replicate 2)	
					Platform Specific Analysis	Nexus Analysis
Affymetrix Arrays						
Affymetrix CytoScan HD	2,600,000 Copy Number + 750,000 SNP	0.88 kb mean probe spacing in RefSeq genes	Affymetrix	Affymetrix Chromosome Analysis Suite	201,241	82,93
Affymetrix Genome-Wide Human SNP Array 6.0	946,000 Copy Number + 906,000 SNP	0.7 kb mean probe spacing	Affymetrix & Mc Carroll et al.	Birdsuite[a]	151[a]	105,112
Agilent Arrays						
Agilent SurePrint G3 Human CGH Microarray, 1×1M (Design ID 021529)	963,029 (60mers)	2.1 kb overall median probe spacing (1.8 kb in Refseq genes)	Service Provider	Agilent Genomics Workbench 7.0	124,133	65,53
Agilent SurePrint G3 Human High Resolution Microarray, 1×1M (Design ID 023642)	963,331 (60mers)	2.6 kb overall median probe spacing, 3 kb average probe spacing	Service Provider	Agilent Genomics Workbench 7.0	155,152	177,171
Agilent SurePrint G3 Human CGH Microarray, 2×400K (Design ID 021850)	411,056 (60mers)	5.3 kb overall median probe spacing (4.6 kb in Refseq genes)	Service Provider	Agilent Genomics Workbench 7.0	58,61	50,48
Agilent SurePrint G3 Human CNV Microarray, 2×400K (Design ID 021365)	442,892 (60mers)	1 kb overall median probe spatial resolution in CNVs	Service Provider	Agilent Genomics Workbench 7.0	365,365	433,489
Agilent SurePrint G3 Human CGH Microarray, 4×180K (Design ID 022060)	170,334 (60mers)	13 kb overall median probe spacing (11 kb in Refseq genes)	Service Provider	Agilent Genomics Workbench 7.0	24,25	30,32
Illumina Arrays						
Illumina HumanOmni5Exome v1	4,641,128	0.63 kb mean, 0.32 kb median probe spacing	Illumina	Illumina GenomeStudio 2011.1 and cnvPartition 3.2.0	426,397	66,62
Illumina HumanOmni5-4v1	4,301,331	0.68 kb mean, 0.36 kb median probe spacing	Illumina	Illumina GenomeStudio 2011.1 and cnvPartition 3.2.0	63,61	56,61

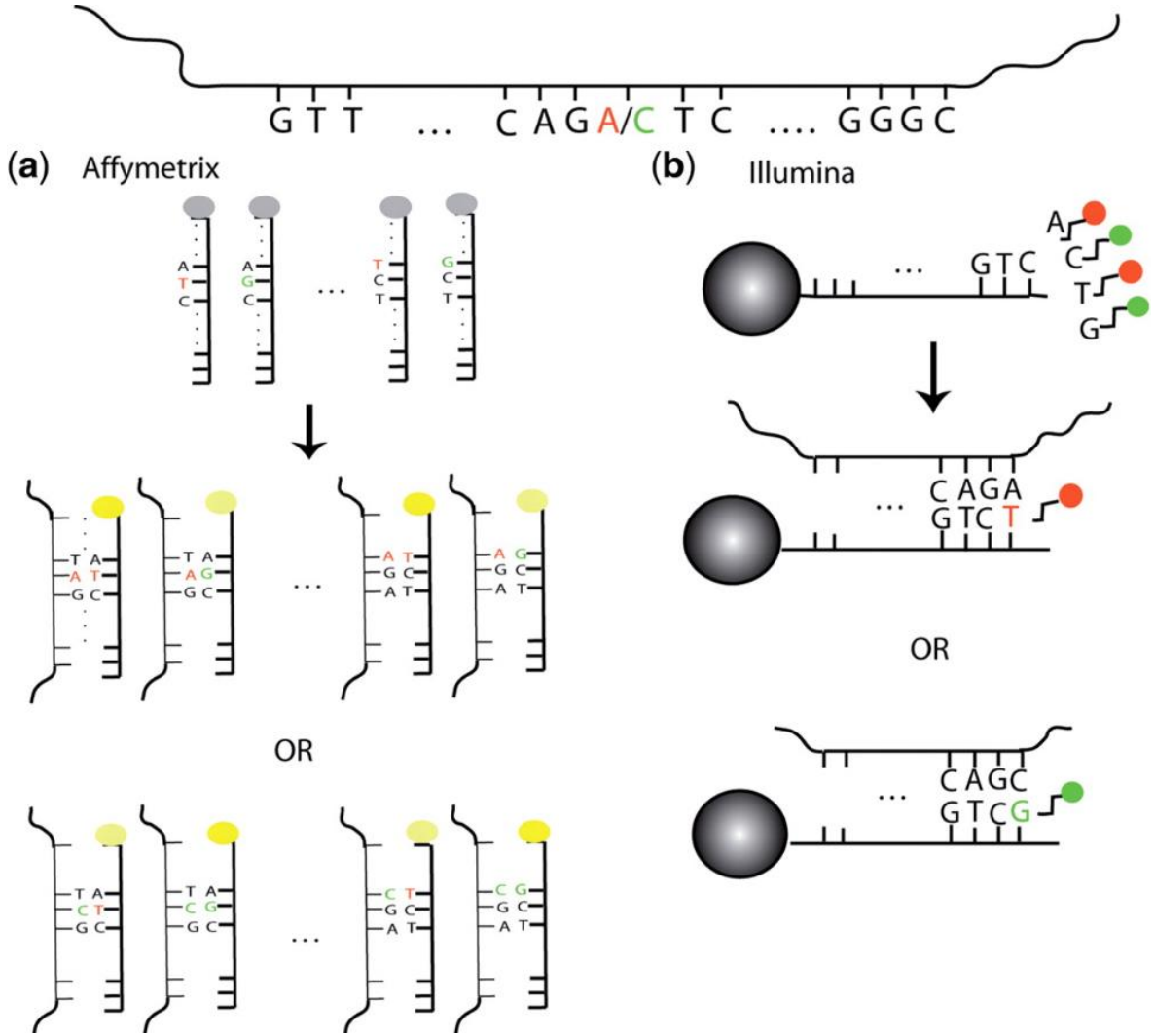
Arrays: basics



The comparative genomic hybridization (CGH) array compares the patient's DNA to control DNA using 2 different fluorescent labels. Labeled control and patient DNA fragments are hybridized to an array containing oligonucleotide DNA sequences from genes throughout the human genome. Each position on the array correlates to a different part of the genome. The relative intensity of the 2 different labels indicates copy-number changes. When only the red label (control DNA) is present, it indicates an absence of patient DNA and therefore a deletion (red stars). When there is more patient than control DNA, the patient label is overrepresented (green circles) and indicates duplication. When there are no copy-number changes, there should be equal amounts of control-labeled and patient-labeled DNA (indicated with blue circles).

A single-nucleotide polymorphism (SNP) array contains small fragments of DNA from the human genome where there are known to be multiple alleles. Each allele is represented on the array and each position on the array corresponds to a genetic locus. DNA from the patient is hybridized to the array. Patients who have the A allele at a specific locus will bind to the A allele on the array. If the patient is homozygous, the sample will bind only to A or B (AA or BB). If the patient is a heterozygote, the sample will be label hybridized to A and B (AB). Copy-number changes are determined by the relative intensity of bound DNA at each allele with a relative decrease in deletions (red bar) and an increase in duplications (green bar). Consanguinity is indicated by a loss of heterozygosity over large spans of DNA.

FIGURE 1



Arrays: basics

Information used:

- Log R Ratio = $\log(R/R_{exp})$
- B Allele Frequency

- The Log R Ratio (LRR) is a normalized measure of the total signal intensity for two alleles of the SNP.
- The B Allele Frequency (BAF) is a normalized measure of the allelic intensity ratio of two alleles.
- The combination of LRR and BAF can be used to infer copy number changes in the genome.
- For example, in the presence of a deletion, there is a decrease in LRR values and a lack of heterozygotes in BAF values; in the presence of a duplication, there is an increase in LRR values and a split of heterozygotes in two clusters

Arrays: basics

CNV type	Possible SNP genotypes	Expected A+B signal	Expected BAF
Normal	*AA*	2	0
	AB	2	0.5
	BB	2	1
Hemizygous loss	*A_*	1	0
	*B_*	1	1
Homozygous loss	*_ _*	0	Undefined

CNV type	Possible SNP genotypes	Expected A+B signal	Expected BAF
Homozygous gain	*AAAA*	4	0
	AAAB	4	0.25
	AABB	4	0.5
	ABBB	4	0.75
	BBBB	4	1
Hemizygous gain	*AAA*	3	0
	AAB	3	0.33
	ABB	3	0.67
	BBB	3	1

Searching for Variants?

Database of Genomic Variants (DGV; http://dgv.tcag.ca/

Database of Genomic Variants

A curated catalogue of human genomic structural variation

Database	Focus	Link
1000 Genomes	Human genetic variation	http://www.1000genomes.org
COSMIC cancer database	Somatic mutations in cancer	http://cancer.sanger.ac.uk/cosmic
dbGAP	Genotype, Phenotype	http://www.ncbi.nlm.nih.gov/gap
dbSNP	SNP	http://www.ncbi.nlm.nih.gov/SNP
HapMap	Haplotype map of human genome	http://hapmap.ncbi.nlm.nih.gov
HGDP	Genetic diversity in humans	http://www.hagsc.org/hgdp
ICGC	Oncogenic mutations	https://icgc.org
PharmGKB	Genetic variation on drug response	https://www.pharmgkb.org
SNPedia	Effect of DNA variations	http://www.snpedia.com/index.php/SNPedia
TCGA	Genetic changes in cancers	http://cancergenome.nih.gov

COSMIC Catalogue Of Somatic Mutations In Cancer

TCGA

International Cancer Genome Consortium

RELATED LINKS

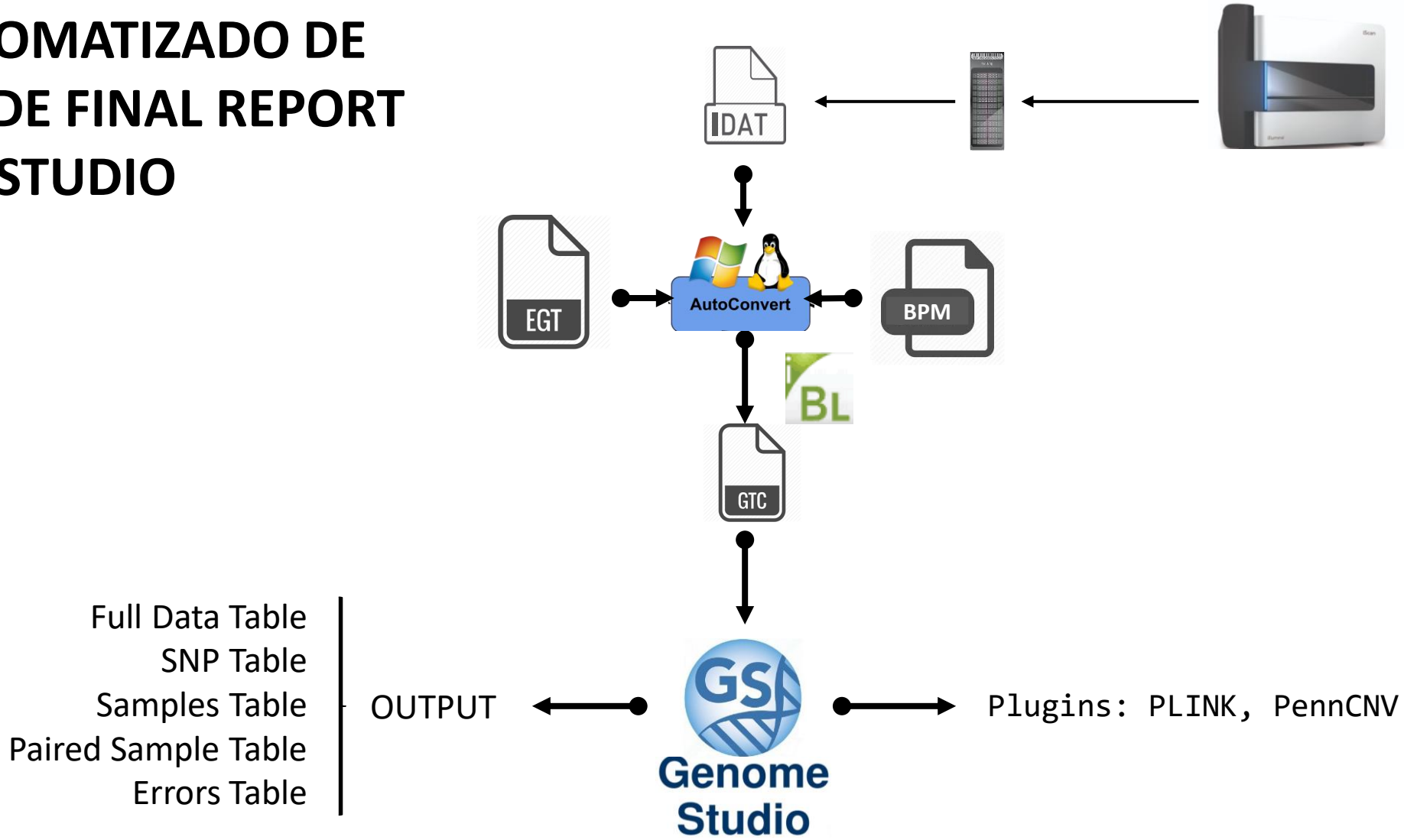
Database of Genomic Variants: http://dgv.tcag.ca/
3D Genome Browser: http://promoter.bx.psu.edu/hi-c/
VISTA Enhancer Browser: https://enhancer.lbl.gov/
DECIPHER: https://decipher.sanger.ac.uk/

The most comprehensive SV database to date, the Database of Genomic Variants (DGV) provides a catalogue of SVs found in the genomes of control individuals from diverse populations. The current version includes >22,300 genomes. dbSNP has >350.000.000 SNP allocated.

Experimental Biologist
Computational Biologist
NO EATING
NO DRINKING
NO BREATHING
NO OPEN TOE SHOES
NO SHORTS
NO MINI SKIRTS
ALWAYS LAB COATS ON
LIFE IS BEAUTIFUL

Number of Markers	Fixed markers: ~ 640,000 Custom marker add-on capacity: Up to 50,000
Input Quantity	200 ng
Description	An economical next-generation genotyping array that enables population-scale genetics, translational research, variant screening studies, and precision medicine research by combining highly optimized multiethnic genome-wide content, curated clinical research variants, and QC markers.
Number of Samples	24 samples per array
Sample Throughput	~2304 samples per week (estimate for 1 iScan System, 1 AutoLoader, 2 Tecan robots, and a 5-day work week)
Assay Type	Infinium HTS
System Compatibility	iScan
Method	Genome-Wide Genotyping Array, High-Throughput Genotyping Array
Variant Class	Copy Number Variants (CNVs), Germline Variants, Insertions-Deletions (indels), Single Nucleotide Polymorphisms (SNPs), Structural Variants
Automation Capability	Automated Array Loader, Liquid Handling Robots
Specialized Sample Types	Blood, Buccal Swabs, Saliva
Nucleic Acid Type	DNA
Technology	Microarray
Species Category	Human

PROCESO AUTOMATIZADO DE GENERACIÓN DE FINAL REPORT CON GENOMESTUDIO



1 EVALUACIÓN DEL DISEÑO DEL CHIP Y CORRECCIÓN DE HEBRAS

BPM → GitHub → PL

StrandScript: evaluation of Illumina genotyping array design and strand correction

BPM → Chipendium → .bim File, .txt File

Detailed match report

The table below shows potential chips matches alongside the total overlap with *your* chip. Chips are recommended based upon the percentage overlap for each chip, alongside the percentage overlap with *your* chip. Overlaps of less than 95% are not recommended.

Chip ID	Most Likely	No. markers	Total marker overlap	% of *this* chip	% of *your* chip

2 CUSTOM-SCRIPTS

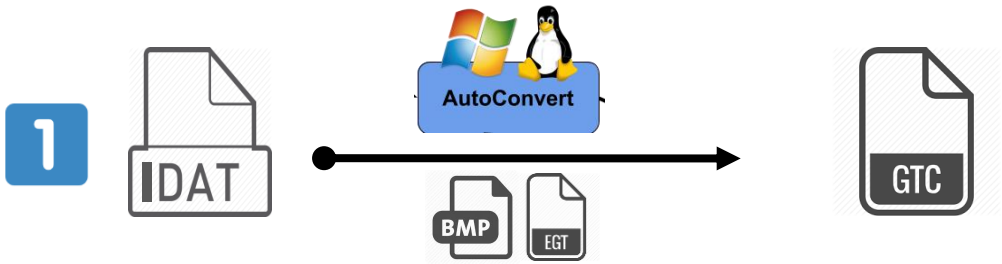
BPM → PL `Mee2Human.pl` Script →

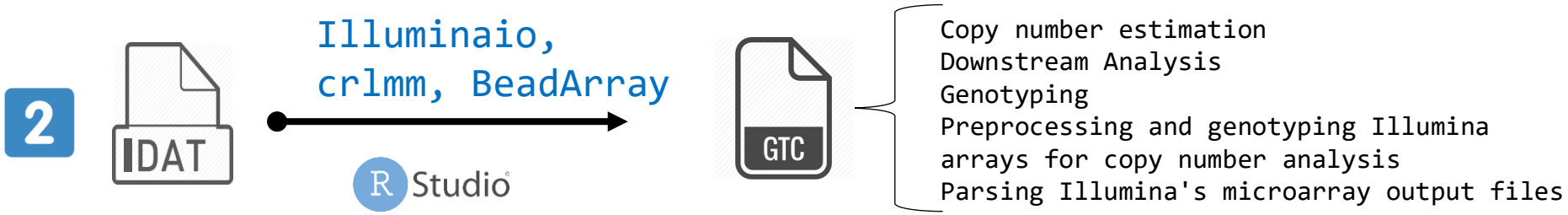
Extrae, entre otros campos, el alelo A y alelo B para cada sonda del manifest.

BPM → `BPM2vcf.py` Script →

Genera un fichero vcf /v=4.1 a partir del manifest

ESTRATEGIAS DE ANÁLISIS A PARTIR DE DATOS CRUDOS (.idat)





ESTRATEGIAS DE ANÁLISIS A PARTIR DE GenTrain/GenCall (.gtc)

1 GTC → gtc_final_report.py locus_summary.py → Output outputLocus → SH → TXT tellmeGen

2 GTC → gtc_to_vcf.py → Convert Illumina GTC files to VCF files

3 GTC → BL Genome Studio → OUTPUT → Plugins: PLINK, PennCNV

Full Data Table
SNP Table
Samples Table
Paired Sample Table
Errors Table

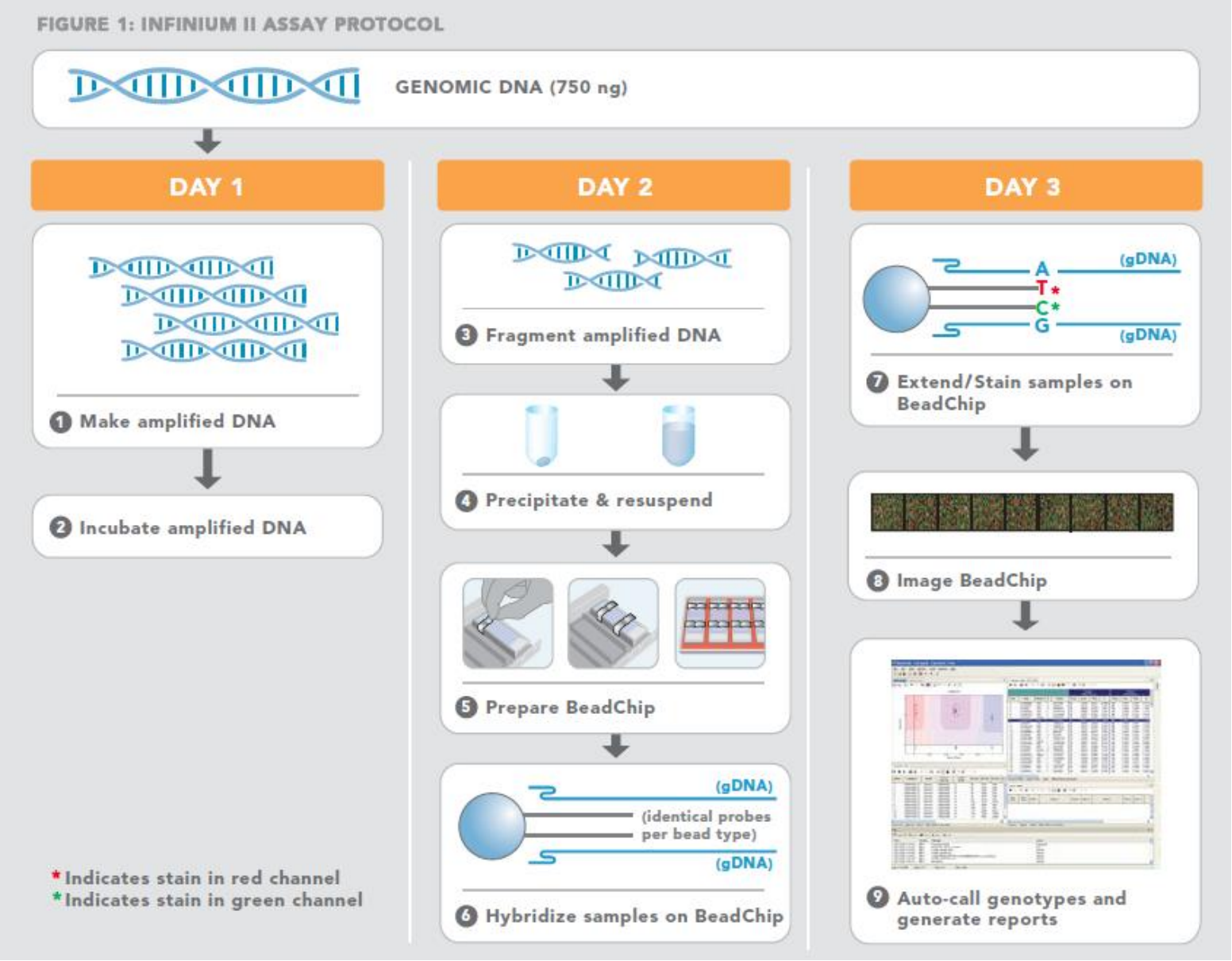


FIGURE 1: INFINIUM II ASSAY PROTOCOL
GENOMIC DNA (750 ng)
DAY 1
DAY 2
DAY 3
1 Make amplified DNA
2 Incubate amplified DNA
3 Fragment amplified DNA
4 Precipitate & resuspend
5 Prepare BeadChip
(gDNA)
(identical probes per bead type)
(gDNA)
6 Hybridize samples on BeadChip
A
T*
C*
G
(gDNA)
(gDNA)
7 Extend/Stain samples on BeadChip
8 Image BeadChip
9 Auto-call genotypes and generate reports
*Indicates stain in red channel
*Indicates stain in green channel

Genome Studio

GenomeStudio® Genotyping Module v2.0 Software Guide

Technical Note: Genotyping

illumina

Infinium® Genotyping Data Analysis

A guide for analyzing Infinium genotyping data using the GenomeStudio® Genotyping Module

PROTOCOL

Illumina human exome genotyping array clustering and quality control

Yan Guo[1], Jing He[2], Shilin Zhao[1], Hui Wu[1], Xue Zhong[1], Quanhu Sheng[1], David C Samuels[3], Yu Shyr[1] & Jirong Long[2]

[1]Center for Quantitative Sciences, Vanderbilt University, Nashville Tennessee, USA. [2]Vanderbilt Epidemiology Center, Vanderbilt University, Nashville Tennessee, USA. [3]Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee, USA. Correspondence should be addressed to Y.G. (yan.guo@vanderbilt.edu).

Strategies for processing and quality control of Illumina genotyping arrays

Shilin Zhao, Wang Jing, David C. Samuels, Quanghu Sheng, Yu Shyr and Yan Guo

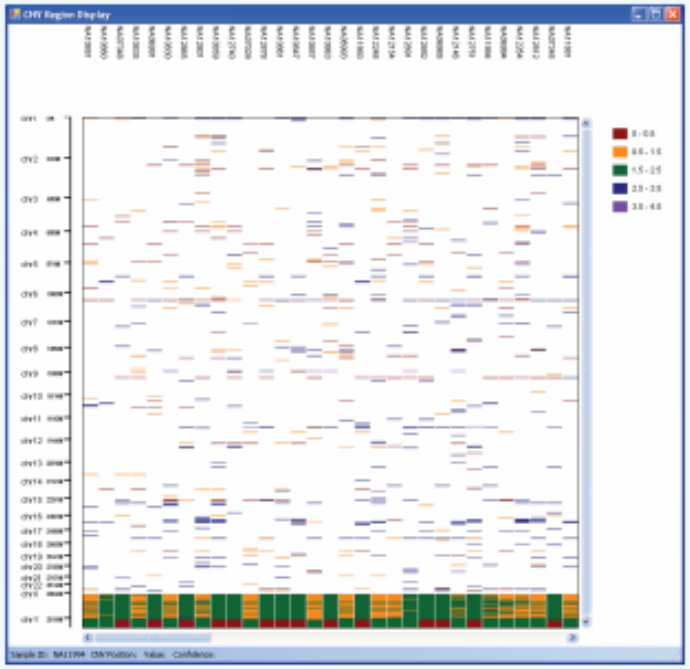
Corresponding author. Yan Guo, Department of Cancer Biology, Vanderbilt University, Nashville, TN 37235, USA. Tel.: 615-936-0816; E-mail: yan.guo@vanderbilt.edu

COPY NUMBER ANALYSIS WORKFLOW WITH GENOMESTUDIO

1 Load sample sheet, *.idat, *.bpm, and *.egt files into GenomeStudio from both cases and controls.

2 Run cnvPartition on sample set. OR Export data to third-party software analysis package via plug-in API.

3 Examine for common aberrations in the CNV Region Display (below) or in a heat map.



4 Display aberrations as bookmarks in the available tools.

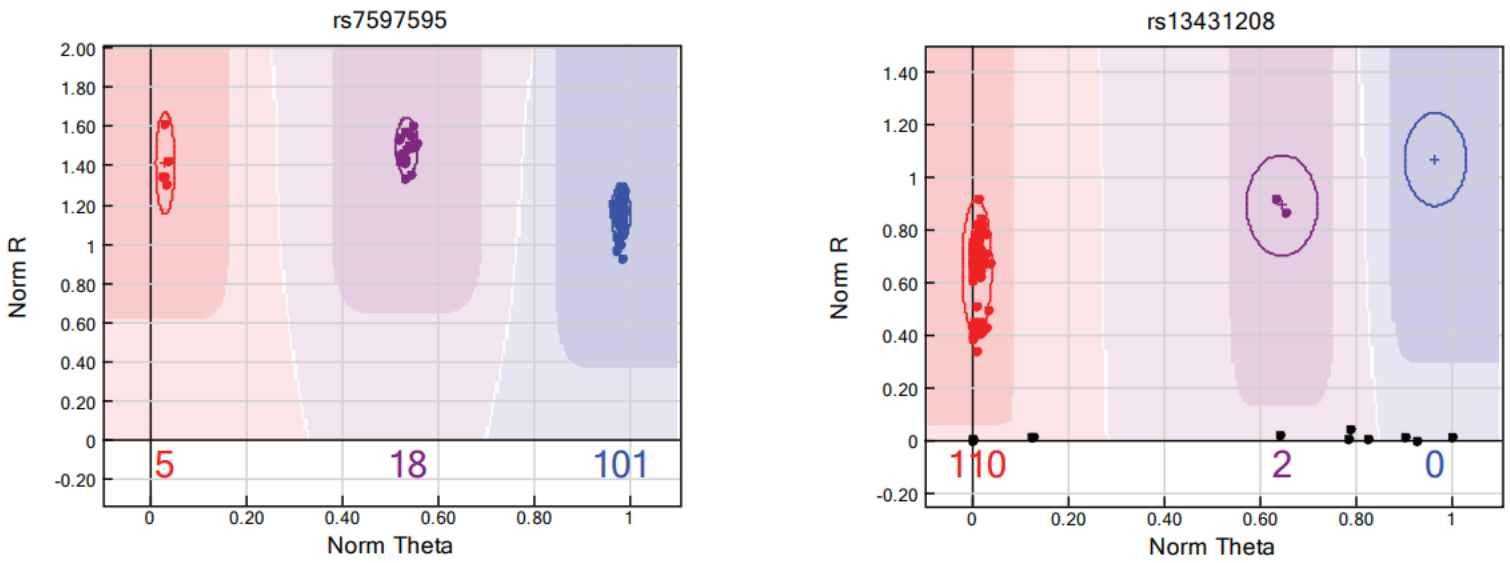
5 Export regions from all samples.

Optional

6 Perform case vs. control analysis of the two sets.

Genome Studio

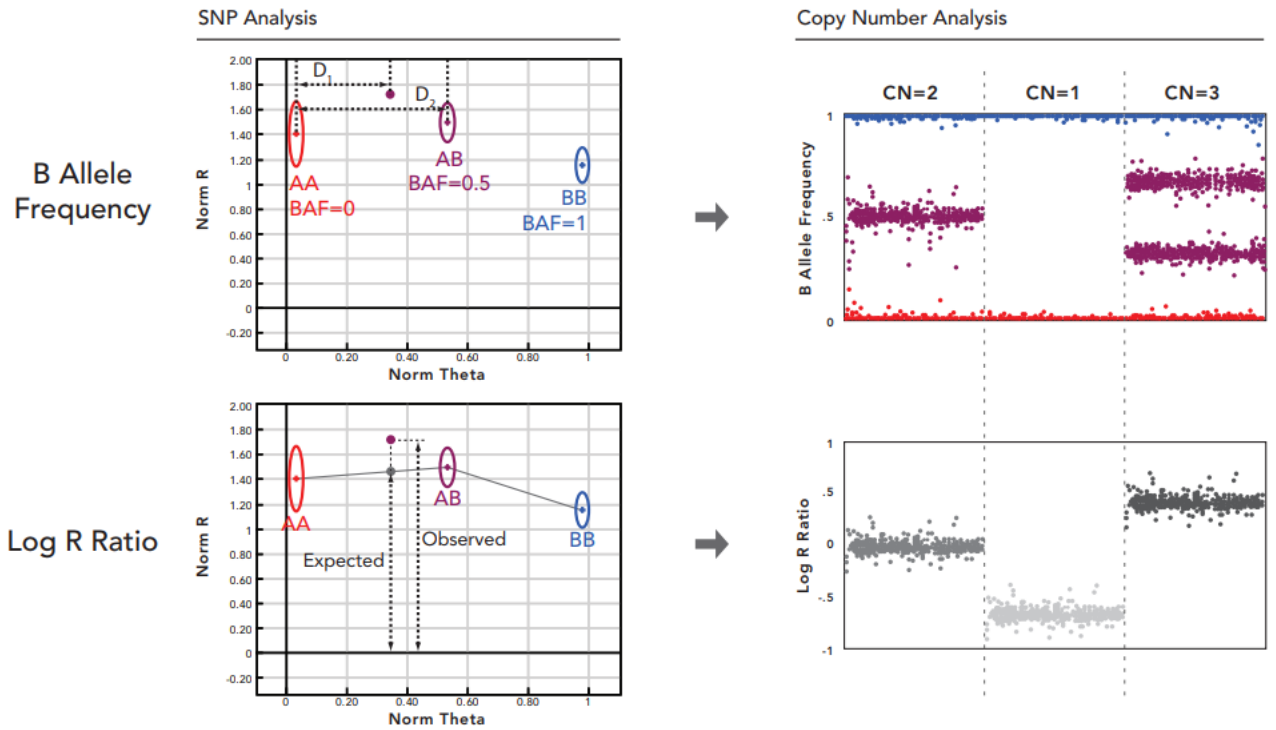
Figure 1: SNP Genoplot Examples



The left panel shows a normal genoplot with samples falling in each of three genotype clusters (red points are AA, purple points are AB, blue points are BB, expected cluster positions are indicated by ellipses). In the genoplot shown in the right panel, the samples represented by black points show a dramatic drop in intensity, which may signify a homozygous deletion in those samples.

Genome Studio

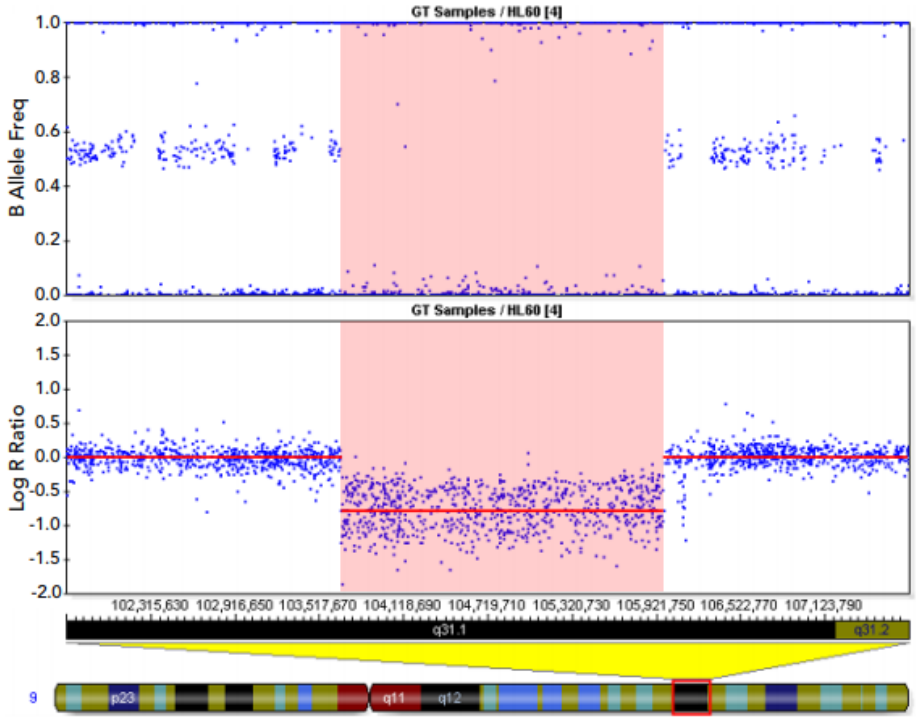
Figure 3: Calculation of Log R Ratio and B Allele frequency



The allelic copy ratio in terms of B allele frequency (BAF) is calculated from the θ value of a sample and the expected cluster positions (ellipses, left top panel). The allele frequency is determined as a linear interpolation in the θ-dimension related to the allele frequency of each cluster (0.0, 0.5, and 1.0). In this example, a data point (purple dot) falling approximately 2/3 of the distance from the AA to the AB cluster (D1/D2 = .66) has an allele frequency of 0.33 (0.66 * 0.5).
The $\log_2$ R ratio is calculated as the ratio between observed normalized intensity of the experimental sample to the expected intensity (left bottom panel). The expected intensity is determined as a linear interpolation as a function of the sample θ (grey line) of the expected cluster positions (ellipses).
These two transformed parameters, B allele frequency and $\log_2$ R ratio, are then plotted along the entire genome for all SNPs on the array (right panel). The plot of these two parameters exhibits diagnostic signature profiles of copy number (example copy numbers 2, 1, and 3 shown) and specific classes of structural variation.
Figure adapted from Peiffer, et. al, 2006.

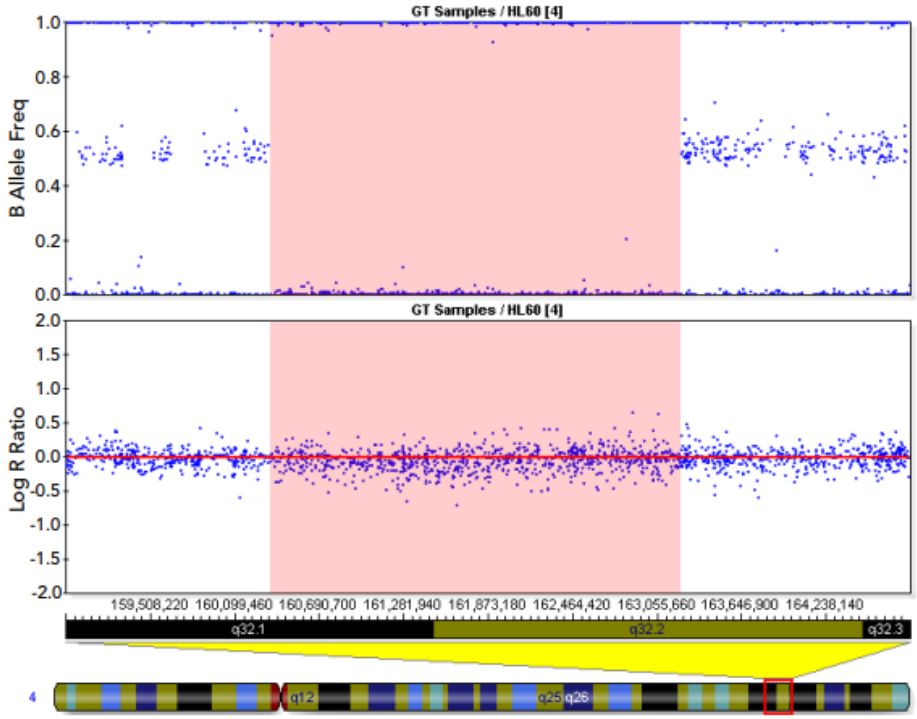
Genome Studio

Figure 5: Hemizygous Deletion



A hemizygous deletion (loss of one copy) is depicted in the ICB as a loss of heterozygotes in the B Allele Freq plot (top) and a loss of signal intensity in the Log R Ratio plot (bottom). In the region of the deletion (shaded), the log R ratio is log2 of 1/2, or -1.

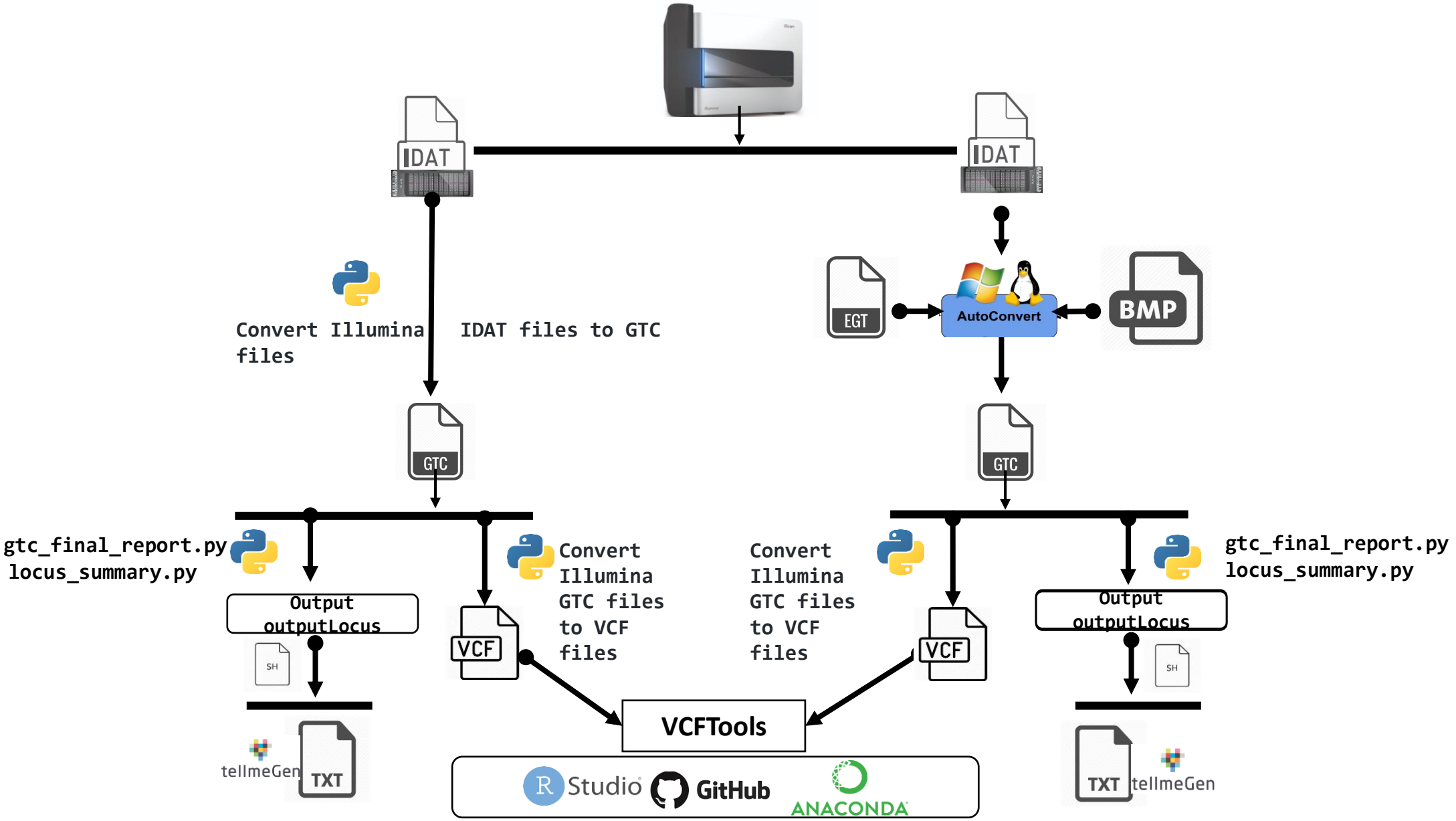
Figure 6: Copy-Neutral LOH



A region of copy-neutral LOH (shaded) is depicted by a loss of heterozygotes in the B allele frequency data but no change in the log R ratio (physical copy number).

- How much copy number variation (CNV) exists between human genomes?
- What is the contribution of copy number variation to genetic disease?
- What is the relative contribution of different mutational mechanisms to CNV?
- What is the genomic impact of CNV on gene expression?
- What role has copy number variation played in recent human evolution?

PART 2: detecting CNV



PART 3: hands on

An overview of the copynumber package

Gro Nilsen

April 30, 2018

Nilsen *et al. BMC Genomics* 2012, **13**:591
http://www.biomedcentral.com/1471-2164/13/591

BMC Genomics

SOFTWARE Open Access

Copynumber: Efficient algorithms for single- and multi-track copy number segmentation

Gro Nilsen[1,2†], Knut Liestøl[1,2†], Peter Van Loo[3,4], Hans Kristian Moen Vollan[5,6,7], Marianne B Eide[2,8], Oscar M Rueda[9], Suet-Feung Chin[9], Roslin Russell[9], Lars O Baumbusch[5], Carlos Caldas[9,10], Anne-Lise Børresen-Dale[5,6] and Ole Christian Lingjærde[1,2,5*]

https://bioconductor.org/packages/release/bioc/vignettes/copynumber/inst/doc/copynumber.pdf

Package ‘copynumber’

June 3, 2018

Title Segmentation of single- and multi-track copy number data by penalized least squares regression.

Version 1.20.0

Author Gro Nilsen, Knut Liestoel and Ole Christian Lingjaerde.

Maintainer Gro Nilsen <gronilse@ifi.uio.no>

Description Penalized least squares regression is applied to fit piecewise constant curves to copy number data to locate genomic regions of constant copy number. Procedures are available for individual segmentation of each sample, joint segmentation of several samples and joint segmentation of the two data tracks from SNP-arrays. Several plotting functions are available for visualization of the data and the segmentation results.

License Artistic-2.0

LazyData yes

Date 2013-04-16

BuildResaveData best

Depends R (>= 2.10), BiocGenerics

Imports S4Vectors, IRanges, GenomicRanges

biocViews aCGH, SNP, CopyNumberVariation, Genetics, Visualization

Copynumber Workflow

Data	Preprocessing	Segmentation	Visualization
Copy number data (+ allele frequencies)	**Outlier handling** *winsorize(...)* **Missing value imputation** *imputeMissing(...)*	**Individual segmentation of one or more samples** *pcf(...)* **Joint segmentation of multiple samples** *multipcf(...)* **Segmentation of SNP-array data** *aspcf(...)*	**Whole-genome plots** *plotHeatmap(...)* *plotGenome(...)* *plotCircle(...)* *plotFreq(...)* **Chromosome plots of data and segments** *plotSample (...)* *plotChrom (...)* *plotAllele(...)* **Diagnostic plot** *plotGamma (...)*

```
1  ##########################################################################
2  #                Msc. PELAYO G. DE LENA RODRiGUEZ                        #
3  #                R version 3.5.3 (2019-03-11)                            #
4  #                  RStudio Version 1.2.1335                              #
5  #          Tutorials for the R/Bioconductor Package copynumber           #
6  #      Instituto de Estudios Celulares y Moleculares - ICM               #
7  #                     MADRID, 16/05/2019                                 #
8  ##########################################################################
9
10
11 # start with a clean slate
12
13 rm(list=ls(all=TRUE))
```

```r
# INSTALL PACKAGE FROM bioconductor

source("https://bioconductor.org/biocLite.R")
biocLite("copynumber")
```

`subsetData` *Retrieve a data subset*

Description

This function returns a subset of copy number data according to the input and the specified chromosomes and/or samples.

Usage

```
subsetData(data, chrom = NULL, sample = NULL, sep="\t", ...)
```

```
# load, inspect and subset dataset
data(lymphoma)
View(lymphoma)
sub.lymphoma <- subsetData(data=lymphoma)
sub.lymphoma[1:10,]
```

winsorize	*Winsorization of copy number data*

Description

Outliers in copy number data are detected and modified using MAD or PCF Winsorization.

Usage

```
winsorize(data, pos.unit = "bp", arms = NULL, method = "mad", tau = 2.5,
          k = 25, gamma = 40, iter = 1, assembly = "hg19", digits = 4,
          return.outliers = FALSE, save.res = FALSE, file.names = NULL,
          verbose = TRUE)
```

```
# use winsorize function to detect outliers
lymph.wins <- winsorize(data=sub.lymphoma,verbose=FALSE)
lymph.wins[1:10,]
```

`winsorize`	*Winsorization of copy number data*

Description

Outliers in copy number data are detected and modified using MAD or PCF Winsorization.

Usage

```
winsorize(data, pos.unit = "bp", arms = NULL, method = "mad", tau = 2.5,
          k = 25, gamma = 40, iter = 1, assembly = "hg19", digits = 4,
          return.outliers = FALSE, save.res = FALSE, file.names = NULL,
          verbose = TRUE)
```

```
# retrieve results
wins.res <- winsorize(data=sub.lymphoma,
                      return.outliers=TRUE,
                      verbose=TRUE)
```

pcf	*Single-sample copy number segmentation.*

Description

Fit a individual piecewise constant segmentation curve to each sample's copy number data.

Usage

```
pcf(data, pos.unit = "bp", arms = NULL, Y = NULL, kmin = 5, gamma = 40,
      normalize = TRUE, fast = TRUE, assembly = "hg19", digits = 4,
      return.est = FALSE, save.res = FALSE, file.names = NULL, verbose = TRUE)
```

```
# segmentation algorithm pcf
single.seg <- pcf(data=lymph.wins,gamma=12,verbose=FALSE)
```

multipcf	*Multi-sample copy number segmentation.*

Description

Joint segmentation resulting in piecewise constant curves with common break points for all samples.

Usage

```
multipcf(data, pos.unit = "bp", arms = NULL, Y = NULL, gamma = 40,
          normalize=TRUE, w=1, fast = TRUE, assembly = "hg19", digits = 4,
          return.est = FALSE, save.res = FALSE, file.names = NULL, verbose
          = TRUE)
```

```
multi.seg <- multipcf(data=lymph.wins,verbose=FALSE)
head(multi.seg)
```

`plotGenome`	*Plot copy number data and/or segmentation results*

Description

Plot copy number data and/or segmentation results for the whole genome.

Usage

```
plotGenome(data = NULL, segments = NULL, pos.unit = "bp", sample = NULL,
           assembly="hg19", winsoutliers = NULL, xaxis = "pos",
           layout = c(1,1), ...)
```

```
plotGenome(data=sub.lymphoma,segments=single.seg,sample=1,cex=3)
```

`plotSample`	*Plot copy number data and/or segmentation results by sample*

Description

Plot copy number data and/or segmentation results for each sample separately with chromosomes in different panels.

Usage

```
plotSample(data = NULL, segments = NULL, pos.unit = "bp", sample = NULL,
            chrom = NULL, assembly = "hg19", winsoutliers = NULL, xaxis =
            "pos", layout = c(1,1), plot.ideo = TRUE, ...)
```

```
plotSample(data=sub.lymphoma,segments=single.seg,layout=c(5,5),sample=1,cex=3)
```

`plotChrom`	*Plot copy number data and/or segmentation results by chromosome*

Description

Plot copy number data and/or segmentation results for each chromosome separately with samples in different panels.

Usage

```
plotChrom(data = NULL, segments = NULL, pos.unit = "bp", sample = NULL,
            chrom = NULL, assembly = "hg19", winsoutliers = NULL, xaxis =
            "pos", layout = c(1,1), plot.ideo = TRUE, ...)
```

```
plotChrom(data=lymph.wins,segments=multi.seg,layout=c(3,1))
```

Usage

```
data(logR)
data(BAF)
```

Format

Two corresponding data sets containing 10000 probes with logR and BAF measurements, respectively, for 2 samples. The two first columns in both data sets contain chromosome numbers and local probe positions (in base pairs), while the subsequent columns contain logR-values and BAF-values in the two data sets, respectively.

```
# load, inspect the dataset
data(logR)
data(BAF)
```

winsorize	*Winsorization of copy number data*

Description

Outliers in copy number data are detected and modified using MAD or PCF Winsorization.

Usage

```
winsorize(data, pos.unit = "bp", arms = NULL, method = "mad", tau = 2.5,
          k = 25, gamma = 40, iter = 1, assembly = "hg19", digits = 4,
          return.outliers = FALSE, save.res = FALSE, file.names = NULL,
          verbose = TRUE)
```

```
# use winsorize function to detect outliers
logR.wins <- winsorize(logR,verbose=FALSE)
```

Index 40

aspcf	*Allele-specific copy number segmentation.*

Description

Joint segmentation of SNP array data resulting in piecewise constant curves with common break points for copy number data and B-allelle frequency data.

Usage

```
aspcf(logR, BAF, pos.unit = "bp", arms = NULL, kmin = 5, gamma = 40,
      baf.thres=c(0.1,0.9), skew = 3, assembly= "hg19", digits = 4,
      return.est = FALSE, save.res = FALSE, file.names=NULL, verbose = TRUE)
```

```
# segmentation algorithm
allele.seg <- aspcf(logR.wins,BAF,verbose=FALSE)
head(allele.seg)
```

plotAllele	*Plot SNP data and/or aspcf segmentation results*

Description

Plot bivariate SNP data and/or aspcf segmentation results for each sample separately with chromosomes in different panels

Usage

```
plotAllele(logR = NULL, BAF = NULL, segments = NULL, pos.unit = "bp",
           sample = NULL, chrom = NULL, assembly="hg19", baf.thres =
           c(0.1,0.9), winsoutliers = NULL, xaxis = "pos", layout = c(1,1),
           plot.ideo = TRUE, ...)
```

```
plotAllele(logR,BAF,allele.seg,sample=1,chrom=c(1:4),layout=c(2,2))
```

`plotFreq` *Plot percentage of samples with an aberration at a genomic position*

Description

Plot the percentage of samples that have an amplification or deletion at a genomic position. Amplifications/deletions correspond to copy number values that are above/below a pre-defined threshold. Frequencies may be plotted over the entire genome or separately for each chromosome.

Usage

```
plotFreq(segments, thres.gain, thres.loss = -thres.gain, pos.unit = "bp",
    chrom = NULL, layout = c(1, 1),...)
```

```
#Run pcf
seg <- pcf(data=lymphoma,gamma=12)
#Plot over entire genome, gain and loss thresholds are 0.1 and -0.1:
plotFreq(segments=seg,thres.gain=0.1)
#Plot by chromosomes, two sets of thresholds:
plotFreq(segments=seg,thres.gain=c(0.1,0.2),
         thres.loss=c(-0.05,-0.1), chrom=c(1:23),
         layout=c(5,5))
```

`plotCircle`	*Plot a circular genome with aberration frequencies and connections between genomic loci added.*

Description

A circular genome is plotted and the percentage of samples that have a gain or a loss at a genomic position is added in the middle of the circle. Gains/losses correspond to copy number values that are above/below a pre-defined threshold. In addition arcs representing some connection between genomic loci may be added.

Usage

```
plotCircle(segments, thres.gain, thres.loss = -thres.gain, pos.unit = "bp",
    freq.colors = c("red","limegreen"), alpha = 1/7, arcs = NULL,
    arc.colors = c("goldenrod1","dodgerblue"), d = 0.3, assembly = "hg19")
```

```
lymphoma.res <- pcf(data=lymphoma,gamma=12,verbose=FALSE)

chr.from <- c(2,12,4)
pos.from <- c(168754669,847879349,121809306)
chr.to <- c(14,21,17)
pos.to <- c(6147539,301955563,12364465)
cl <- c(1,1,2)
arcs <- cbind(chr.from, pos.from, chr.to, pos.to,cl)
```

`plotCircle`	*Plot a circular genome with aberration frequencies and connections between genomic loci added.*

Description

A circular genome is plotted and the percentage of samples that have a gain or a loss at a genomic position is added in the middle of the circle. Gains/losses correspond to copy number values that are above/below a pre-defined threshold. In addition arcs representing some connection between genomic loci may be added.

Usage

```
plotCircle(segments, thres.gain, thres.loss = -thres.gain, pos.unit = "bp",
    freq.colors = c("red","limegreen"), alpha = 1/7, arcs = NULL,
    arc.colors = c("goldenrod1","dodgerblue"), d = 0.3, assembly = "hg19")
```

```
plotCircle(segments=lymphoma.res,thres.gain=0.2,arcs=arcs)
```

`plotHeatmap`	*Plot copy number heatmap*

Description

Heatmap reflecting the magnitude of estimated copy numbers relative to some pre-defined limits. Estimates may be obtained using `pcf` or `multipcf`, and results may be visualized over the entire genome or by chromosomes.

Usage

```
plotHeatmap(segments, upper.lim, lower.lim = -upper.lim, pos.unit = "bp",
    chrom = NULL, layout = c(1, 1),...)
```

```
plotHeatmap(segments=lymphoma.res,upper.lim=0.3)
```

`plotAberration` *Plot areas with copy number aberrations*

Description

Create plots reflecting the location of aberrated segments. Results may be visualized over the entire genome or by chromosomes.

Usage

```
plotAberration(segments, thres.gain, thres.loss = -thres.gain, pos.unit = "bp",
    chrom = NULL, layout = c(1, 1),...)
```

```
plotAberration(segments=lymphoma.res,thres.gain=0.2)
```

`micma`	*Subset of 244K aCGH data*

Description

A subset of the 244K MicMa data set containing copy number measurements for six samples on chromosome 17.

Usage

```
data(micma)
```

Format

Data frame containing 7658 probes with log2-ratio copy numbers for 6 samples on chromosome 17. The first column contains the chromosome numbers, the second gives the local probe positions (in base pairs), while the subsequent columns contain the copy number measurements for the individual samples.

```
# MODEL PARAMETERS: gamma
data(micma)
```

`plotGamma` *Plot segmentation results for several values of gamma*

Description

Data for one sample on one chromosome is segmented by `pcf` for 10 values of gamma, and results are visualized in a multi-grid plot.

Usage

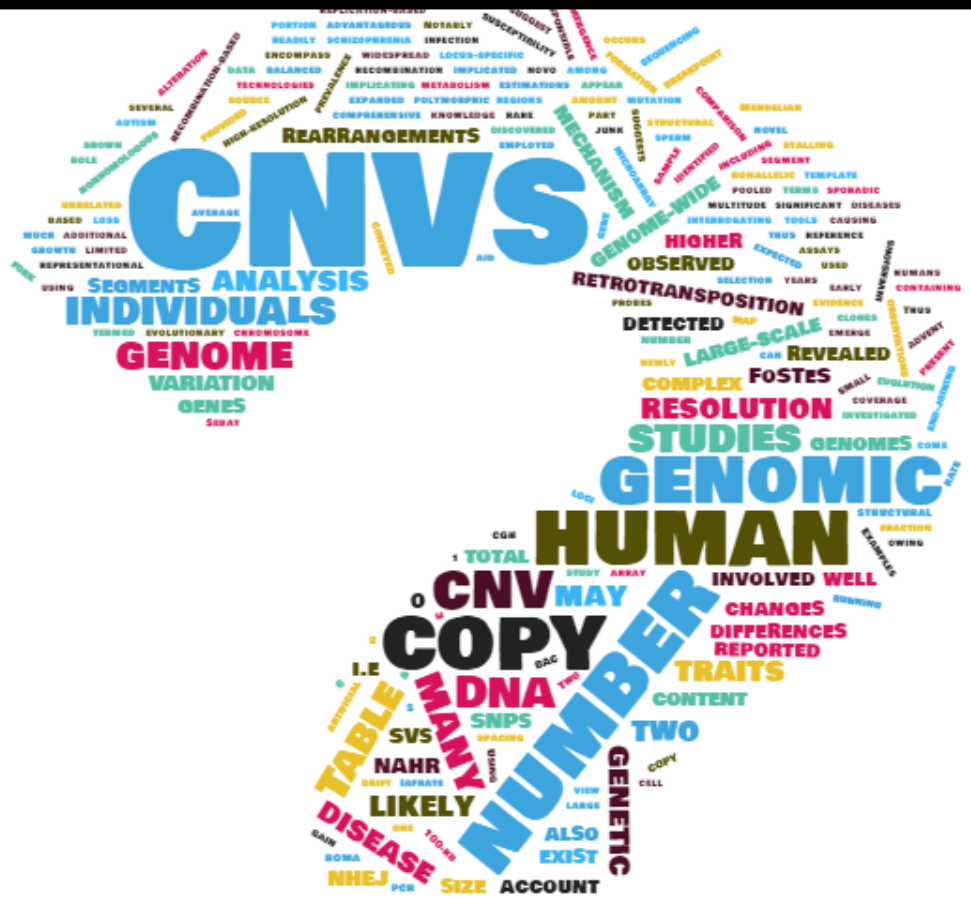
```
plotGamma(data, pos.unit = "bp", gammaRange = c(10,100), dowins = TRUE,
          sample = 1, chrom = 1, cv = FALSE, K = 5, cex = 2, col = "grey",
          seg.col="red", ...)
```

```
plotGamma(micma,chrom=17,cex=3)
```

INSTITUTO DE ESTUDIOS CELULARES Y MOLECULARES

ATCGCGCGCATTGCTAGATCAATGCTAGCTAGAGCTGCACGGGATGCGCTGCTAGCTGACTGACTTCGTAGCTAGTCGATCGTAGTCGATGCTATGCATCGCGCGCATTGCTAGATCAATGCTAGCTAGAGCTGCACGGGATGCGCTGCTAGCTGACTGACTTCGTAGCTAGTCGATCGTAGTCGATGCTATGCGCGGCGCCGGCGCGCCG

Unidad de Formación en Energías Renovables, Medioambiente y Biotecnología · ciberonc Centro de Investigación Biomédica en Red Cáncer · R Studio · Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas



ATCGCGCGCATTGCTAGATCAATGCTAGCTAGAGCTGCACGGGATGCGCTGCTAGCTGACTGACTTCGTAGCTAGTCGATCGTAGTCGATGCTATGCATCGCGCGCATTGCTAGATCAATGCTAGCTAGAGCTGCACGGGATGCGCTGCTAGCTGACTGACTTCGTAGCTAGTCGATCGTAGTCGATGCTATGCGCGGCGCCGGCGCGCCG