bioinformatics.ca

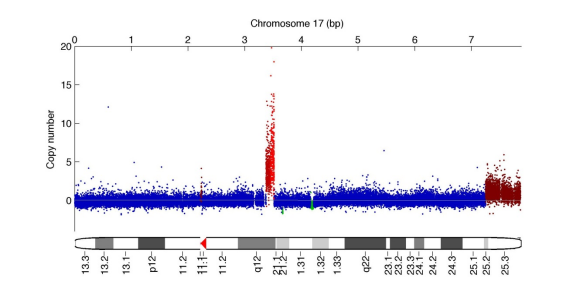
Canadian Bioinformatics Workshops

www.bioinformatics.ca

Lab Module 5
Copy Number Alterations

Sohrab Shah, Fong Chun Chan
Bioinformatics for Cancer Genomics
May 25-29, 2015

bioinformatics.ca



BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

Many slides adapted from Gavin Ha, Andrew Roth, and Andrew McPherson

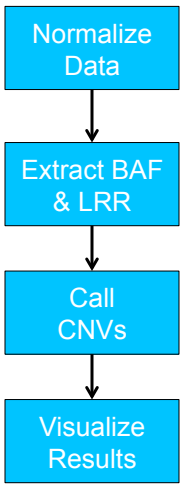
Lab Module 5: Learning Objectives

1. Understand the basic workflow for identifying CNVs
2. Apply methods to identify CNAs in array and sequencing data
3. Visualize and interpret the results of CNA analyses

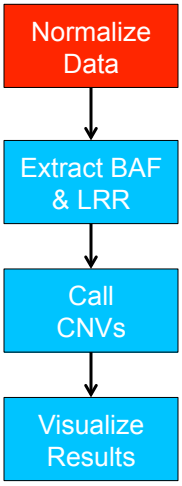
Getting Started

- The “Installation” wiki page shows installation of tools used in this module
- The “PrepData” wiki page details where we obtained the data
- The “Lab” wiki page shows the steps taken in this lab
- We will work through the “Lab” wiki page. You can use the “Installation” and “PrepData” wiki pages as references if interested

Analysis Of CNAs using Arrays



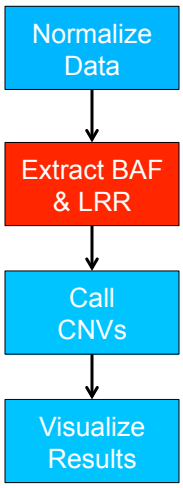
Step 1: Normalize Data



Step 1: Normalize Data

- Because DNA does not hybridize with the same efficiency in all experiments, some normalization must be done
- Methods for normalization depend on the platform (e.g. Affymetrix SNP 6.0, Illumina)
- Software (Affymetrix SNP 6.0)
  - Affymetrix Power Tools http://www.affymetrix.com/support/developer/powertools/index.affx
  - Using R the AROMA package can be used http://www.aroma-project.org/
- For this tutorial we will use the Affymetrix Power Tools
  - Comprehensive description of the entire normalization process can be found here: http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html

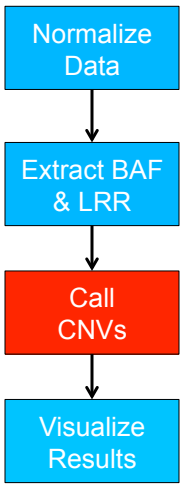
Step 2: Extract BAF and LRR



Step 2: Extract BAF and LRR

- SNP genotyping arrays can measure how many copies of the A and B alleles are present
- Most tools prefer to work with B allele frequencies (BAF) and Log R Ratio (LRR):
  - BAF = B / (A+B)
  - LRR = log(A + B) – log(x)
    - x is the intensity we would expect from a probe at normal copy number
- We need to convert the normalized outputs from the arrays to these values
- For this tutorial we used normalize_affy_geno_cluster.pl supplied as part of the PennCNV package http://www.openbioinformatics.org/penncnv/

Step 3: Call CNVs



Step 3: Call CNVs

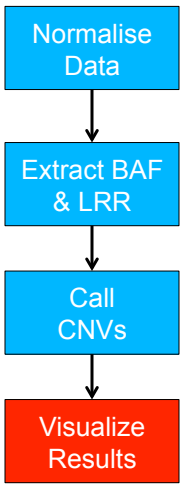
- Calling CNVs converts the BAF and LRR measurements across the many probes (~1.8 million for Affymetrix SNP 6.0) to predictions of copy number segments

- Calling CNVs in tumour genomes is challenging due to:
  - Normal contamination
  - Germline polymorphisms
  - Genomic heterogeneity of tumour cells
  - Ploidy i.e. baseline LRR

Step 3: Call CNVs

- Software (a few examples):
  - OncoSNP (https://sites.google.com/site/oncosnp/)
  - PICNIC (http://www.sanger.ac.uk/genetics/CGP/Software/PICNIC/)
  - ASCAT (http://heim.ifi.uio.no/bioinf/Projects/ASCAT/)
  - HAPSEG (https://confluence.broadinstitute.org/display/CGATools/HAPSEG)
- For this tutorial we will use OncoSNP:
  - It is well documented
  - Handles all the confounding factors in a statistically sound framework

Step 4: Visualize Results



Step 4: Visualize Results

- OncoSNP results can be found in the module package:
  - content/results/oncosnp
- The plot files HCC1143.*.ps.gz provide a nice summary figure of the data

CNV Analysis using Sequencing Data

- Conceptually, CNV analysis for sequencing and array data follows the same workflow
- The main difference is that we start from aligned read data (BAM) instead of raw array data (CEL, etc)
- A new issue that arises with sequencing data is mappability:
  - Some regions of the genome are easier to map reads to than other regions

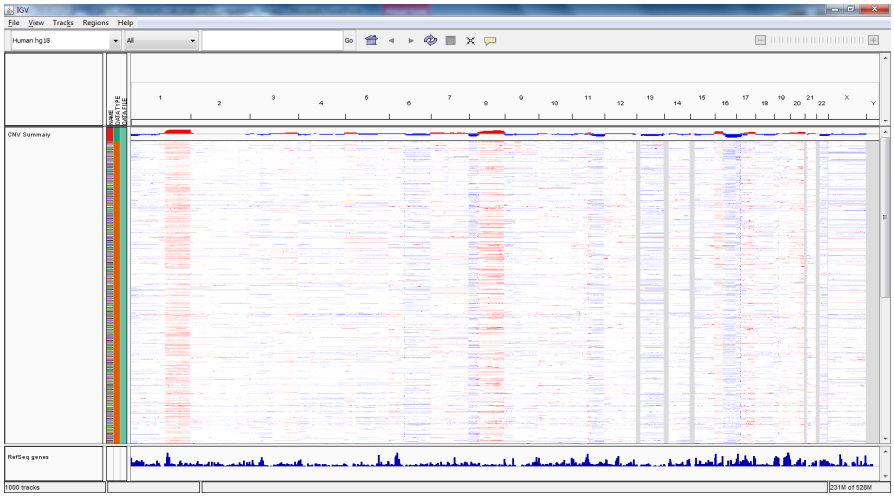
CNV Analysis using Sequencing Data

- Software (a few examples)
  - OncoSNP-Seq (https://sites.google.com/site/oncosnpseq/)
  - HMMCopy/TITAN (http://compbio.bccrc.ca/software/titan)
  - SomitCA (http://www.bioconductor.org/packages/release/bioc/html/SomatiCA.html)
  - …
- Today we will HMMCopy/TITAN because they are published and relatively mature

Browse Segment Data in IGV

1. Download METABRIC_DatasetI997.seg from the wiki
   - 997 breast cancer copy number alterations predicted from SNP6 arrays (Curtis C*, Shah SP*, et al., Nature, 2012)
2. Open IGV
3. Switch reference genome to “Human hg18”
4. File -> Load from File…
5. Tracks -> Fit Data to Window

Browse Segment Data in IGV



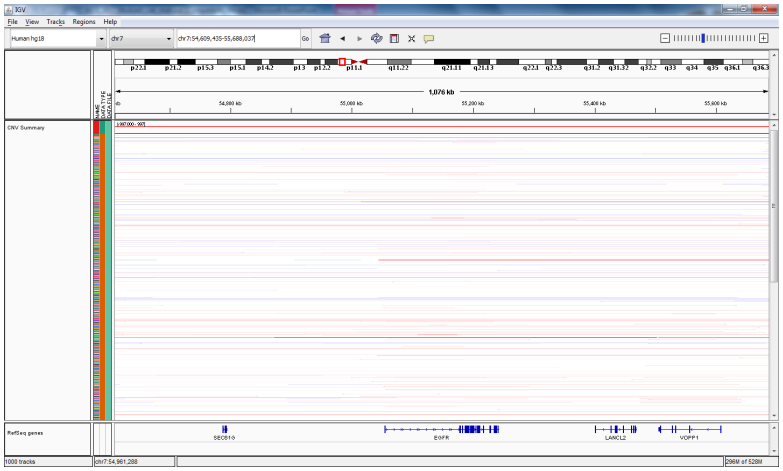
Example of Somatic High-Level Amp

Examples of EGFR somatic high-level amp

In the Search bar, type the gene:

EGFR

Press Enter, then click zoom out "-" twice.



Example of Germline CNVs

Examples of germline CNVs

In the Search bar, type the gene:

GSTM1

Note the alternating gains and losses.
Note the length of the markers