solution technology

SNP array

Pros

SNP array data offers many advantages in terms of cost, throughput, turn around time and processing. But the key advantage of SNP arrays is the use of SNP BAF in addtion to LRR which really increase the CNV detection sensitivity and a better identification of the event type.

Cons

SNP array are limited to detecting copy-number differences of sequences present in the reference assembly used to design the probes, provide no information on the location of duplicated copies. A major limitation of SNP array is the size range of alterations detected. In order to provide good CNV call a minimum nbumber of probes is required which allows only the detection of large event and disable the resolution of breakpoints at the single-base-pair level.

NGS

Pros

The most important benefits of NGS technologies are a genome-wide analysis without a priori information, the specificity and linear dynamic range response of NGS data offer many advantages for estimation of copy number. Additionally NGS allows fine detection of small event and several methods of analysis coming from the CGH arrays domain are available.

Cons

NGS data are limited by the cost, the turn around time and the processing complexity. Moreover the major limitations of NGS approach for CNV are the use of uniform reads distribution assumption (False) and the inherente noise in the data introduced by the quality of the reference sequence during the alignment step.

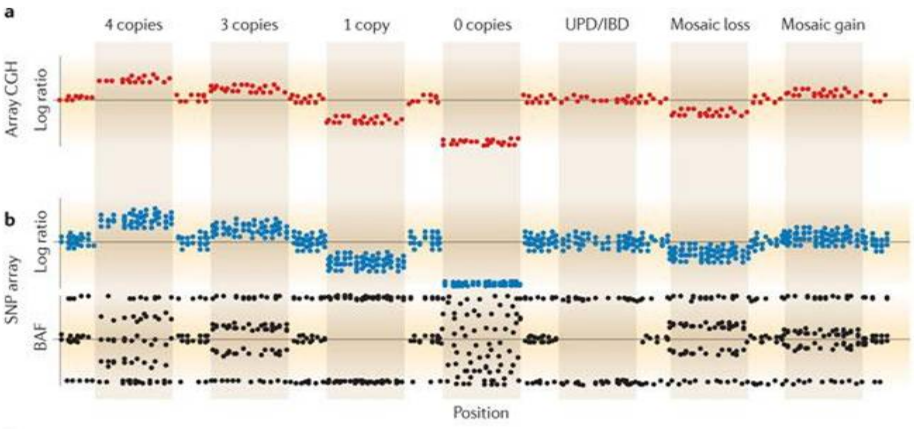

Figure 1: NGS (CGH) vs. SNP-array data representation

NGS (CGH) vs. SNP-array data representation

Conclusion

When studying CNV, using a combined approach of SNParray and NGS is actually the best alternative.

solution NgsAnalysisSummary

here are the steps to do:

1. Sequence trimming
2. Genome alignment
3. Alignment refinement
4. Bin creation and count
5. Computing LRR
6. LRR signal smoothing (optional)
7. LRR signal segmentation
8. CNV calling from segments

solution BinCount

bincounter parameters:

Parameter	explanation	Sugested va
–minMapQ	Filter out reads a mapping quality lower or equal at	

Parameter	explanation	Sugested va
–gc	Compute GC% per window	*not neccess*
–ref	Path to the indexed Reference Genome	*not neccess*
–refbam	Path to the germline sample bam	../alignment/normal/normal.sorted.dup.b
–bam	Path to the tumor sample bam	../alignment/tumor/tumor.sorted.dup.b
–norm	unit for count normalization	
–windows	bin size	1

solution BinCount 2

As we are using partial genome data and the count has been made on the entire genome we don't expect to find coverage in these regions

solution cancerChallenge

Tumor ploidy It is well known that many types of tumors frequently have genomic aberrations involving gain or loss of whole or large parts of chromosomes. Thus, the average ploidy or total genomic content of tumor cells cannot be assumed to be 2N.

Conventional microarray copy number analysis is based on comparing the probe intensities to those of a set of diploid reference samples. This works well for detecting aberrations in diploid non-cancer samples as the normalized intensity of copy number two should coincide for query and reference data. However, many individual tumors have such extensive genomic aberrations that the assumption that the query cells have a genomic content of 2N on average is severely violated.

Tumor heterogeneity Tumor samples are a mix of cancer cells and genetically normal cells due to samping step.

The proportion of tumor cells can vary considerably, complicating the analysis since the measured signal from any locus will be a combined signal from both tumor and non-tumor cells. If the proportion of tumor cells is too low, aberrations will remain undetected.

Tumor cell heterogeneity Copy number aberrations in tumor cells may arise several times throughout tumor development, and may give rise to different subclones. The extent of the proliferative advantage and the time between occurrence of the aberration and tumor collection influence the proportion of tumor cells with each aberration [15].

The average copy numbers of heterogenic genomic regions may be non-integer. Long non-integer regions may severely disturb a model-based copy number analysis as they do not fit the pre-determined relationship between signal and copy number.

solution SnpAnalysisSummary

SNP data analysis for CNV detection

1. Generate LRR
2. Generate BAF
3. Probe filtering
4. segmenetation of LRR and BAF signal
5. CNV calling from segments