Detecting copy-number variants from Next generation sequencing data
Gunjan Gala and David Gresham
Center for Genomics and Systems Biology, Department of Biology, New York University

Background:
Genetic differences between organisms are often encoded in the form of structural variations, such as insertions, deletions and duplications, etc. Evolutionarily related genomes may not only exhibit single-nucleotide polymorphisms (SNPs) and short insertion or deletion (indel) variations, but also structural variations (SV) consisting of larger chromosomal deletions, insertions, and rearrangements. The impact of these structural variants (SVs) has been demonstrated in a wide range of applications including disease association studies, cancer genomics, and evolutionary studies. Next Generation Sequencing (NGS) provides opportunities for genome-wide SV assays with higher resolution and larger categories of variations than the conventional microarray-based methods. Different methods for detecting CNVs are split-read(SR), read-pair(RP), read-depth(RD) or read-count(RC) and assembly (AS) based methods as shown in figure 1. Out of the many tools based on these methods, we select a representative for each method i.e. Pindel(SR), Breakdancer(RP), Cnvnator(RD) and Lumpy(SR+RP) and compare their efficiency in terms of false discovery rate (FDR) and sensitivity.

Future work:
1. Study performance of these tools on different datasets eg reads generated from PacBio, Third generation sequencing technology (TGS), etc and/or datasets with lower coverage eg 10,50x, etc
2. These tools did not detect insertions effectively. So for detecting insertions, we would try an assembly based method eg Novelseq.
3. Use tools like picard or samtools to remove duplicates from PCR products prior to structural variants detection.

References:
3. Abyzov A et al. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21(6):974-84.

Acknowledgements:
We would like to thank members of Gresham Lab for their feedbacks and Sidharth Jain for his help with data simulations.