**AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision.**

A Abyzov, M Gerstein (2011) Bioinformatics 27: 595-603.

The paper proposes a new sequence alignment algorithm that allows the identification of breakpoints of SV. The evaluation of potential SV breakpoint is done using the alignment scoring function. Given any two sequences two scoring alignment matrices are generated. One matrix “walks” the alignment from 5' to 3' while the other scans in opposite direction (3' to 5'). The scoring matrices are based mainly on Smith-Waterman algorithm or Needelman-Wunsch algorithm in the case of cDNA. Maximum cell score indicates the starting point for finding the optimum local alignment. Final best score is the summation of the best scores of 5'-3' and 3'-5' matrices defining the optimal alignment. Method tested primary on sequences containing a single SV, mainly a deletion.

**Mapping copy number variation by population-scale genome sequencing.**

RE Mills et al. 1000 Genomes Project (2011) Nature 470: 59-65.

The paper describes the identification and characterisation of SVs from 1000 Genome project with focus on deletions. A variety of methods (pair-end mapping (PR), read-depth analysis (RD), split-read analysis (SR), sequence assembly (AS) and a combined  PR and RD (PD)) are used to identify the SVs. The SVs are characterised based on genotype using GenomStrip and alleles frequency. Results show that PR, RD and PD are the best identification methods. Average SV size is 8kb, 4x smaller than CGH-based analysis. 16k new SVs were identified based on comparison to personal genome and database of structural variation. SVs were classified based on their formation mechanism using BreakSeq. NH is the dominant deletion mechanism while MEI is the preferred insertion mechanism. A series of “preferred” SV formation mechanisms are identified on chromosome “hot-spots” locations.

**Detection of copy number variation from array intensity and sequencing read depth using a stepwise Bayesian model.**

ZD Zhang, MB Gerstein (2010) BMC Bioinformatics 11: 539.

Based on the step function like pattern of genomic signal of CNV data generated from read-depth sequencing a new CNV identification method using Bayesian functions is created.

The existence of a CNV in the real data set is defined as a step function (eg. If a genomic signal data point from is part of CNV, the function is non-zero and equal to the signal amplitude, otherwise the function is 0). The CNVs in the real data set are characterised by the total number of CNVs (N), a trio defining the location (*s)*, width (*w)* and amplitude of signal (*a*) of each CNV, and the signal noise (s). The real data signal is fitted using the parameters obtained from Markov chain Monte Carlo (MCMC) simulation. Gibbs sampling is used to sample the CNV parameters (*s,w,a*). The model selection using a Bayes factor is applied to the identification of *N*. The method was tested successfully on data from array-CGH samples from Glioblastoma Multiforme.

The main disadvantage is the computational demanding MCMC simulation.

**Personal genome sequencing: current approaches and challenges.**

M Snyder, J Du, M Gerstein (2010) Genes Dev 24: 423-31

The paper gives a general overview of the current state of art of genome sequencing and analysis

It covers and discusses the methods used for identification and characterisation of SNPs and SVs. The commonly used methods to estimate the accuracy and sensitivity of SNPs, Indels, SVs and new DNA sequences are discussed. The linear performance metric from Anchor X Prize Competition is analysed. The present day sequencing technologies are listed with highlight on pair-end sequencing. A step-by-step description of genome resequencing strategy is given. **OBS:** Really nice flowchart for sequencing personal genomes.