

COMPARATIVE ANALYSIS OF ALGORITHMS EMPLOYED IN DETERMINATION OF CNV LOCUS AND FREQUENCY

Karthika Periyathambi, Stanford University

Abstract—Copy Number Variants are a new parameter to measure the rate of mutations that in turn are responsible for pathological effects and susceptibility to diseases. This paper is an introduction to the different techniques to estimate CNV location and frequency computationally and comparison of their performances. Furthermore, the paper highlights the type of mutations and elaborates on the importance of CNV estimation. Also, the effect on known diseases and the impact of CNV variations on evolution has been investigated.

Index Terms—CNV (Copy Number Variants), SNP (Single Nucleotide Polymorphism), array based CGH (Comparative Genomic Hybridization, ROC curve based residual, HMM, Penn CNV, Quanti CNV, Bayes Factor, Poisson.

I. INTRODUCTION

COPY NUMBER VARIANTS (CNV) refer to insertions, deletions and other structural variants in gene sequences of size greater than 1KB. Copy number variants (CNVs) underlie many aspects of human phenotypic diversity and provide the raw material for gene duplication and gene family expansion. With increasing interest in understanding phenotype of individuals from their genetic build, structural variation has become a prominent factor. Smaller variations with respect to allele are referred as SNP variations and have been studied in detail. But, over the years, researcher have found that affected individuals have variations that span over wider sequences and hence the rise of structural variant term. Structural variant is the umbrella term comprised of all sorts of genomic variations with DNA segments greater than 1 KB. These may be quantitative (insertions/duplications/deletions), or can be orientation (inversions) or positional (translocations) or non-homogeneous recombination as described in the later sections. Copy Number Polymorphism (CNP) refers to CNV occurring in more than 1 % of the population.

The paper has been divided into five sections. The following section explains types of mutations and CNV in depth. The Section III illustrates the importance of estimating CNV variation rate and loci information; while subsequent ones highlight CNV in special reference to genetic diseases and evolution trends. The Section VI presents a detailed view of the most frequently employed algorithms to calculate CNV locus and frequency along with their associated strengths and weaknesses. The last section gives a conclusion into the

different parameters involved in the choice of these algorithms along with the performance of the chosen few.

II. TYPES OF VARIATIONS

According to Database of Genomic Variants [<http://projects.tcag.ca/variation/>], 3213,401 base pairs (0.11%) of the genomic variation is from SNPs while 40,568,593 bp variations (1.35%) are from CNV.

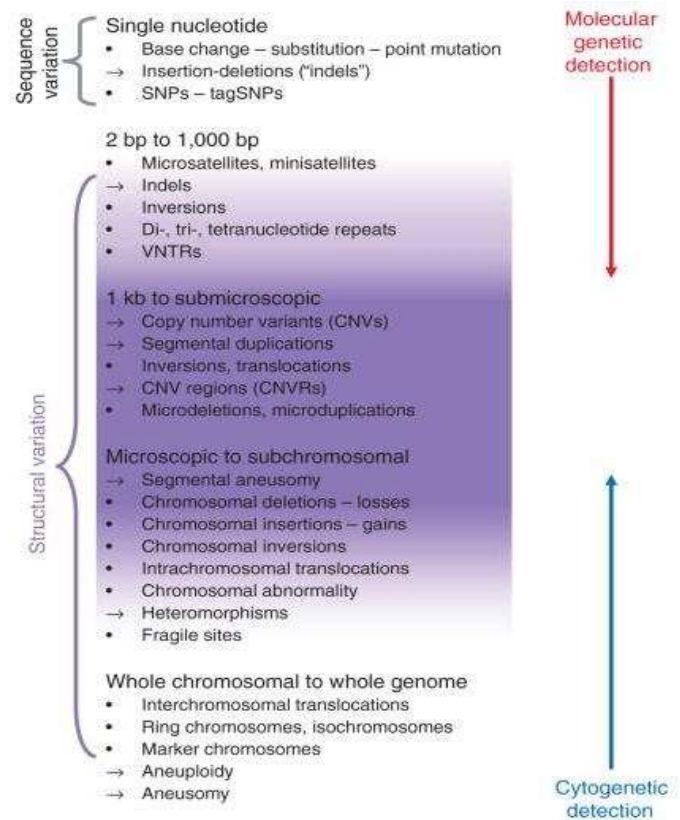


Figure 1

Thus it is clear that average of variations of CNV are about 3-4 times that of SNPs, thus clearly highlighting the importance of studying CNV variations [Iafate *et al.*, 2004 Sebat *et al.*, 2004][6].

Mutations are mostly classified as per the following table:

Type of mutation	Definition	Example
Missense	Codon and hence amino acid is altered	Sickle Cell Anemia A \Rightarrow T at the 17 th nucleotide
Nonsense	Translation prematurely stopped	CAG \Rightarrow TAG in cystic fibrosis
Silent	No change in product	-
Splice-site	Alteration in enzymatic machinery controlling removal of intron	-

Table 1

The above table clarifies the information for SNPs while CNVs have a similar classification as discussed below. The alteration of CNV especially duplications are restricted to a small area of proximity referred to as the CNV locus and CNV region for multiplex arrangements respectively. The CNVs mostly exhibit (i) insertion, (ii) deletion, (iii) duplication either in proximity, homozygous allele or entirely new gene.

About a quarter of these CNVs are flanked by or associated with segmental duplications (SD). SDs are duplicated blocks of genomic DNA typically ranging in size from 1–200 kb, often containing sequence features such as high-copy repeats and gene sequences with intron-exon structure. Over the past decade a large number of both intra- and interchromosomal segmental duplications have been observed.

Insertion and deletion are obvious from their names while duplication can result in one of the sister chromatid to have an extra section while the other one to have it deleted. An illustration for the case of aldosterone steroid has been indicated below:

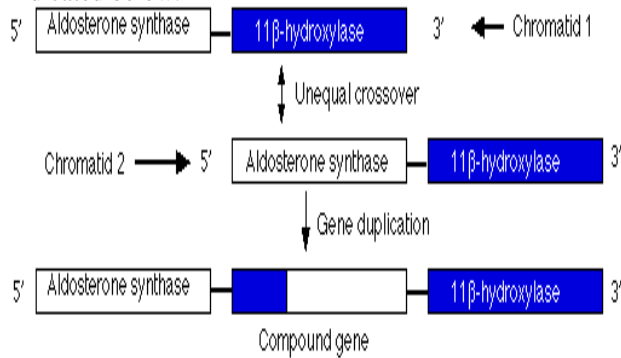


Figure 2

Due to the presence of extra promoter at 5' end, this gene can be expressed more strongly leading to high blood pressure and increased probability of death from stroke. On the other hand, translocations are between different homologous chromosome as in the case of Burkitt's lymphoma.

III. CNV FREQUENCY AND LOCUS ESTIMATION ALGORITHM

Discussing the rate of mutations independent of classifications between CNV and SNP, it can be approximated as:

$$\text{Mutation Rate} = \frac{\text{Total Number of Mutations}}{(\# \text{ Exp}) * (\# \text{ BP}) * \text{avg}(\# \text{ Gens})}$$

,where, # = Number Of
Gens = Generations
BP = Base Pairs
Exp = Experiments

This formula yields to a rough estimate of 2.1×10^{-8} mutations [7] per base pair per generation for Homo sapiens, around 10^{-8} mutations per base pair per generation for Drosophila melanogaster and 7×10^{-9} mutations per base pair per generation for Arabidopsis thaliana (green plant).

CNV Frequency:

Estimating the CNV frequency is important for comprehending the linkage between variants and the phenotypes: asthma, cancer, diabetes and other diseases. Many studies employing computational methods report that, overall, the frequencies of most CNVs are likely to be low [e.g. 95% of CNVs reported by Pinto *et al.* (2007) have frequencies of <2%]. Unlike the mutation frequency estimation, where each mutation is considered independent Poisson distribution, the CNV case is handled differently.

The accuracy of CNV boundaries derived from SNP arrays is influenced by multiple factors such as the robustness of the statistical method, batch effects, population stratification and differences between experiments. The following parts discuss the common approaches employed for locus and frequency estimation.

IV. CNV AS AN EVOLUTIONARY TOOL

Copy Number Variations can draft a relation between genes of different species; while closely related species have lesser CNV differences, the widely separated ones have more CNV differences. Thus, based on CNV statistic matching, evolutionary relations can be derived as illustrated in Figure 3

Human and chimpanzee CNVs [30] occur in orthologous genomic regions far more often than expected by chance and are strongly associated with the presence of highly homologous intrachromosomal segmental duplications.

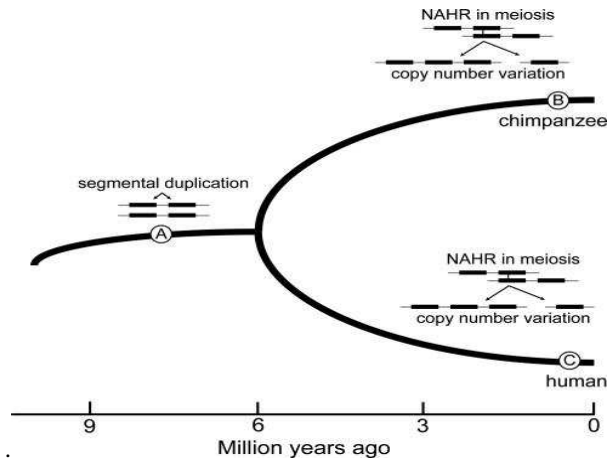


Figure 3

By adapting population genetic analyses for use with copy number data, one can identify functional categories of genes that evolved under purifying or positive selection for copy number change.

V. CNV: THE SECRET KEY TO UNDERSTANDING GENETIC DISEASES

The major reason for researching on CNV arises from the interest in understanding the root cause of genetic diseases. Whole genome approaches have changed the shift from “phenotype first” to “genotype first”. Each of the CNV duplications, deletions, non-homogeneous recombinations have been found to be related to genetic diseases, especially by frequency matching. Segmental duplications mediate diseases owing to more recurrent rearrangements arising from duplicated segments. The following table [2] and figure [4] illustrates the same.

Trait	Rearrangement type	Distance between repeats (kb)	Repeat length (bp)
Color blindness	DEL	0	39 000
α -Thalassemia	DEL	3.7 or 4.2	4000
Growth hormone deficiency	DEL	6.7	2200
Debrisoquine sensitivity	DEL	9.3	2800
Hunter mucopolysaccharidosis	INV	20	3000
Glucocorticoid-remediable aldosteronism	DUP	45	10 000
Hemophilia A	INV	500	9500
CMT1A/HNPP	DUP/DEL	1500	24 011
X-linked ichthyosis	DEL	1900	20 000
Williams syndrome	DEL	~2000	>30 000
Smith-Magenis syndrome/dup(17)(p11.2)	DEL/DUP	~5000	>200 000

Abbreviations: DEL, deletion; DUP, duplication; INV, inversion.

Figure 4 [28]

The unbalanced translocations role has been investigated for germline CNVs (Stranger et al., 2007) and in colon cancer (Camps et al., 2008) and on expression of miRNAs (Zhang et al., 2006). Balanced chromosome rearrangements (translocations, inversions) lead to formation of oncogenic [25] fusion genes (Mitelman et al., 2007) and have been

reported in leukemia, lymphoma and carcinomas related to prostate and lung cancer (Heim and Mitelman, 2008). While reference [26] discusses how del(17)p11.2 leads to Smith-Magenis syndrome; reference [27] discusses mental retardations caused by three interstitial overlapping 17q21.31 microdeletions.

Locus	Del Or dup	Coordinates (Build 36) and size of critical region	Associated Phenotypes	Possible Candidate genes
3q29	Del Dup	197.4-198.9	moderate MR, microcephaly, mild dysmorphic features Duplication: mild to moderate MR	PAK2, DLG3
10q22-q23	del	Chr10: 81.12–89.07 Mb 7.95 Mb	deletion carriers have cognitive and behavioral abnormalities learning disabilities, speech and language delay, mild developmental delays	<i>NRG3</i> <i>GRID1</i> <i>BMPRI</i> <i>ASNCG</i> <i>GLUD1</i>
15q13.3	Del dup	Chr15: 28.7–30.2 Mb 1.5 Mb	Del: mild to severe MR, mild dysmorphism, digital abnormalities, autism; schizophrenia; IGE Dup: mild to moderate delays, has not been reported in schizophrenia or IGE	<i>CHRNA7</i>
1q21.1	Del Dup	chr1: 145.0–146.35 Mb 1.35 Mb	Del: variable phenotypes mild to severe MR, microcephaly, occasional congenital heart disease; enrichment of the deletion in schizophrenia Dup: macrocephaly, mild to moderate delays, autistic features; unlike the deletion, has not been seen in schizophrenia	<i>GJA5</i> , <i>GJA8</i> , <i>HYDIN2</i>
1q21.1	del	chr1: 144.10–144.60 Mb 500 kb	TAR syndrome: hypomegakaryocytic thrombocytopenia, upper extremity abnormalities ranging from bilateral absent	<i>PIAS3</i> , <i>Lix1L</i>

			radii to phocomelia; normal intellect	
15q24	del	Chr15: 72.2– 73.8 Mb 1.8 Mb	Mild to moderate MR, high anterior hairline, downslanting PF, long philtrum, digital abnormalities, genital abnormalities, loose connective tissue	<i>MAN2C</i> <i>1</i> , <i>CYP11</i> <i>A1</i> , <i>STRA6</i>
16p13. 11	Del dup	chr16:15. 4–16.4 Mb 1Mb	Deletion: MR, autism, brain abnormalities Duplication: autism, MR; decreased penetrance	<i>NDE1</i> , <i>NTAN</i>

Table 2 [24]

[Genomic hotspot rearrangements and their associated phenotype]

VI. CNV LOCUS AND FREQUENCY DETECTION TECHNIQUES

The most commonly used algorithms and techniques have been discussed in this section along with a detailed perspective of their merits and demerits.

A) METHOD I: Graph Theory

Each of the markers are considered as vertexes in the graph theory with their connecting sequences as the edges [3].

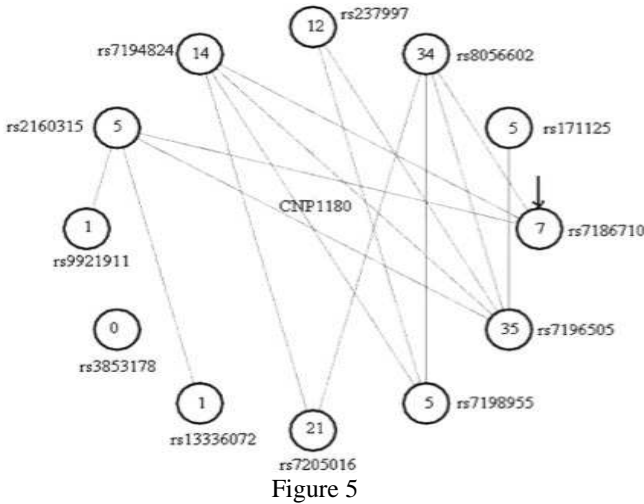


Figure 5

Each marker is assumed to be a likely breakpoint and Poisson process applied to the number of breakpoints summed over all individuals in a sample, with the hypothesis of breakpoints being randomly distributed.

$$P(\text{\#breakpoints at marker } m \geq k) = \sum_{i \geq k} e^{-\lambda} \lambda^i / i!$$

where $\lambda = B/M$, B is the total number of breakpoints, M is the total number of markers and k is the observed number of

breakpoints at the marker under consideration. This gives the frequency estimate of the CNV region with further analysis for estimating the probability of a locus having CNV at its location. Let c_1, \dots, c_l be the \log_2 ratio values at the markers within the CNV, of length l , and x_1, \dots, x_N be the \log_2 ratio values at the markers within a segment of length N , adjacent and to the left of the CNV.

$$t^2 = \frac{\frac{N-l}{N+1} \cdot (\bar{x} - \bar{c})^2}{\frac{1}{N+l-2} (\sum_i (x_i - \bar{x})^2 + \sum_i (c_i - \bar{c})^2)}$$

Under the null hypothesis of equal means, t^2 follows an F -distribution with $(1, N+l-2)$ degrees of freedom. For each individual, the individual is considered as CNV segment if :

$$(\bar{x}_{\text{left}} - \bar{c}) \cdot (\bar{x}_{\text{right}} - \bar{c}) \geq 0 \quad (1)$$

$$P\text{-value } t^2(\text{left}) < \frac{0.05}{\text{\#individuals}} \quad (2)$$

$$P\text{-value } t^2(\text{right}) < \frac{0.05}{\text{\#individuals}} \quad (3)$$

Merits: This graphical technique is an innovative two step process that gives a really close estimate for majority of cases. The two stages ensure the stringency need in the first step and the loose cases to be covered in the second stage.

Demerits: Though a new and innovative technique, the graphical approach has some faulty assumptions :

- For common CNVs (with frequencies > 1%), this method yields lower values owing to the stringent criteria in Step 1.
- False positives are common owing to the difficulty in detecting true CNVs with lower signal to noise ratio.
- This method does not work very well for non-homologous recombination.

B) METHOD II: Bayesian Model for Probes

Reference [1] proposes a dynamic algorithm with Bayesian mathematics to analyze genomic data from multiple sources like oligonucleotide array, bacterial artificial chromosome array and array CGH to minimize noise per probe and severity of chromosomal aberrations. The algorithm employs priorless maximum posteriori estimator and dynamic programming implementation to facilitate discovery of genes and important markers, especially for inherited genetic diseases.

Gaussian distribution with mean μ_r and standard deviation σ_r is assumed for both deviated probes and regular probes (not affected by cancer related chromosomal aberrations) with probability p_r . The mutations are modelled as a Poisson process with parameter $p_b N$, where N is the length of the genome (i.e., total number of probes). The probes along the genome are subdivided into k non overlapping intervals. Probes belonging to a particular interval are assumed to have a similar evolutionary history of duplication and deletion events, and therefore have similar copy-number distributions.

The algorithm has two phases: the first a Poisson distribution to model the number of intervals with Poisson parameter $p_b N$.

The second component is a sequence of Bernoulli trials, one for each probe with probability p_r that a given probe is regular. Combining these factors, the prior distribution becomes

$$Pr(I_N) = e^{-p_r N} \frac{(p_r N)^k}{k!} p_r^{\#regular} (1 - p_r)^{\#deviated}$$

where $\#regular$ is the number of regular probes with the “regular” copy-number distribution and $\#deviated$ is the number of remaining probes in the interval structure I_N . The probabilities can be combined by taking the product of Gaussians:

$$Pr(\mathbf{x}|I_N) = \prod_{i=1}^n \phi(x_i, \mu_j, \sigma^2) \quad 2$$

where the i th probe is covered by the j th interval of the interval structure I_N and μ_j is the mean of the corresponding Gaussian distribution. Dynamic programming minimizes the negative posterior log likelihood function and helps in extending the interval from $I = \langle i_1, \mu_1, \dots, i_k, \mu_k \rangle$, to $I' = \langle i_1, \mu_1, \dots, i_{k+1}, \mu_{k+1} \rangle$, where $i_{k+1} > i_k$ as illustrated below:

$$\begin{aligned} -\log L(I') = & -\log L(I) + \frac{1}{2\sigma^2} \sum_{j=i_k+1}^{i_{k+1}} (x_j - \mu_{k+1})^2 \\ & -\log(p_r N) + \log(k+1) \\ & + \frac{i_{k+1} - i_k}{2} \log(2\pi\sigma^2) - (i_{k+1} - i_k) \\ & \cdot [\mathbb{I}_{k \in regular} \log p_r + \mathbb{I}_{k \in deviated} \log(1 - p_r)] \end{aligned}$$

Merits: This highly precise mathematical technique has few advantages over common approaches.

- Unlike the existent global threshold approaches, this algorithm is not affected by the presence of noise and correlations.
- Unlike the HMM (Hidden Markov Model), the algorithm is not dependent on the exclusive dataset. HMM are sensitive to topology and need normalized dataset.

Demerits: Though this algorithm overcomes the weakness of the existing techniques, this technique is highly computation intensive and is not straightforward. Every step and prediction needs lot of calculations that cannot be hand waived easily, thus increasing the complexity of the algorithm.

C) METHOD III: Bayesian CPCM

Bayesian Analysis can be combined with the knowledge of position of biomarkers and scan statistics as explained in Reference [2] to detect the positions of CNV. Modelling independent Poisson process for the distance between biomarkers, the model follows the steps mentioned below to implement the (Bayesian approach to compound Poisson change-point model):

1. For a chromosome having potentially just one aberration region, equations discussed in paper for

posterior probability and locus \widehat{V} identification can be used.

2. For multiple aberration regions on a chromosome or genome, J sliding windows with sizes from 12 to 35 are chosen such that each window contains exactly one aberration. If w_1, w_2, \dots, w_J , denotes the J windows, $\sum_{i=1}^J w_i$ equals the total number of observations on the chromosome.
3. The number of subintervals ℓ_j with lengths ℓ_j and the number of biomarkers, m_i , in each subinterval are estimated
4. The posterior probabilities for each length is calculated and the locus identified for posterior probabilities greater than 0.5.
5. The identified change positions \widehat{V} are converted to the actual biomarker position $S_{\widehat{V}} = \sum_{i=1}^{\widehat{V}} t_i$, where $S_{\widehat{V}}$ as the position on the chromosome at which the CNV has changed.
6. Steps 3-5 are repeated for varying values of j till convergence is achieved.

Merits: The Bayesian approach has the following distinct advantages:

- It uses both the biomarker positions (distances) and log intensity ratios in its prediction.
- It characterizes the posterior probability of the loci being a CNV which in turns facilitates easy judgment as to if a CNV exists at a locus by using the posterior probability together with their biological knowledge.

Demerits: The window size is not optimally chosen and loosely based on the assumption of eventual convergence. This leads to extraneous computations and probability of false predictions.

D) METHOD IV: SW ARRAY

This method [9] adapts a new approach towards optimizing array CGH analyses using Smith Waterman algorithm to identify genomic regions with CNV spanning multiple probes. SW is applicable based on the visualization of intensity log ratios along the genome as one-dimensional series of continuously distributed scores. Contiguous sequences of predominantly high values in this series may indicate polysomic regions. Conversely, sequences of predominantly low values may indicate deletions, and can be found by changing the sign of the data. The method proceeds as follows:

1. Threshold value t_0 is subtracted from the log ratios, ensuring that the mean of the adjusted scores $X(p)$ is negative.
2. The score of a segment of consecutive probes is the sum of the corresponding adjusted log ratios. The score of the segment from p to q inclusive is

$$T(p, q) = \sum_{i=p}^q X(i).$$

- High-scoring 'islands' are identified using the Smith–Waterman algorithm. A locally high-scoring segment or island is defined to be a positive-scoring segment whose score cannot be increased by shrinking or expanding the segment boundaries. Let $S(p)$ be the score of the island ending at coordinate p , and $B(p)$ to be the coordinate of the beginning of the island. Let $S(0) = 0$, and for $p > 0$; from SW algorithm

$$S(p) = \begin{cases} S(p-1) + X(p) & \text{if } S(p-1) + X(p) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$B(p) = \begin{cases} B(p-1) & \text{if } S(p) > 0 \\ p & \text{otherwise} \end{cases}$$

- The boundaries $\{B(p_{\max}), p_{\max}\}$ and score $S(p_{\max})$ of the overall maximum-scoring island are output by the algorithm.
- In order to identify all islands, the segment corresponding to the maximum-scoring island is replaced by a sequence of zeroes and the algorithm repeated until no positive-scoring islands are detected.

Merits: This algorithm is another unique way of dynamic programming for detecting CNV locus:

- The approach is unique in offering both a nonparametric segmentation procedure and a nonparametric test of significance.
- It is scalable and well-suited to high resolution whole genome array CGH studies that use array probes derived from large insert clones as well as PCR products and oligonucleotides.
- The computational method makes no distributional assumptions about the data to identify putative copy-number changes and determines their statistical significance.
- Adopting SW approach helps in reducing false positives, false negatives complications by polymorphisms/benign variants

Demerits: Though an innovative approach, the algorithm suffers from the following weaknesses:

- Negative-scoring loci can occur inside an island, thus allowing for occasional false positive or negative signals
- The level of resolution depends on the sequence length and spatial density of the arrayed probes. But, increased resolution leads to increased likelihood of identifying imbalances that are due to very small regions representing phenotypically benign variants or polymorphisms.

E) METHOD V: Maximum Likelihood Estimation

Short Tandem Repeats (STRs) are repetitive stretches of DNA made of short sequence motifs (2–6 bp), are very common in eukaryotic genomes, and are highly mutable, with changes in repeat count occurring with much higher mutation rates compared to other polymorphisms thus leading to allelic polymorphism. This method [15] adapts a new approach of assigning binary values to the Poisson parameters.

Let $t(T)$ be the total time of all branches of the phylogenetic tree T , then, the number of mutations on this tree in STR i in total time $t(T)$ is distributed $\text{Poisson}(\lambda_i t(T))$. If T_1, \dots, T_K represent the K Hg terminal subtrees of T , whose total time length of all the branches, t_1, \dots, t_k and inner structure are not known. Thus, m_{ik} , the number of mutations of STR i , in Hg k , is distributed $\text{Poisson}(\lambda_i t_k)$. The m_{ik} 's from the internal structure are not observed and hence can't be directly counted and hence it is tough to formulate the total log-likelihood of the data and estimate the parameters, using Poisson regression, through the usual ML framework, instead the state (number of subunit repeats) of STR i in all samples (leaves) of Hg k is observed.



Figure 6

- (A) The full phylogenetic tree, including the internal Hg phylogenies, which we assume we do not observe. (B) Schematic of the Hg view of a phylogenetic tree.

Instead of observing the Poisson mutation counts m_{ik} :

$$b_{ik} = \begin{cases} 1 & \text{If } m_{ik} = 0 \\ 0 & \text{If } m_{ik} > 0 \end{cases}$$

The binary variables are observed, which are distributed as $b_{ik} \sim \text{Bernoulli}(\exp(-\lambda_i t_k))$. If two different states indicate that at least one mutation occurred, three different states indicate the presence of at least two mutations, etc. The number of mutation events of STR i in Hg k , $m_{ik} \sim \text{Poisson}(\lambda_i t_k)$, hence the probabilities in each case is:

$$\begin{aligned} P(m_{ik} = 0) &= \exp(-\lambda_i t_k) \\ P(m_{ik} > 0) &= 1 - \exp(-\lambda_i t_k) \\ P(m_{ik} > 1) &= 1 - \exp(-\lambda_i t_k) - \lambda_i t_k \exp(-\lambda_i t_k) \\ &\vdots \\ P(m_{ik} > n) &= 1 - \sum_{j=0}^n (\lambda_i t_k)^j \exp(-\lambda_i t_k) / j! \end{aligned}$$

Let y_{ik} be the observed number of states of STR i in Hg k . The log-likelihood of the data y is formulated as:

$$\begin{aligned} \ell(y, \lambda, t) &= \sum_{i,k} [I_{\{(y_{ik}-1)=0\}} \log(P(m_{ik}=0)) \\ &\quad + \sum_{j=1}^8 I_{\{(y_{ik}-1)=j\}} \log(P(m_{ik} > j-1))]. \end{aligned}$$

Merits: This algorithm is an innovative way to employ simplistic procedure for calculating STR and hence extend it to CNV. The rate estimates depend only on polymorphisms

observed, which in turn depend on the total branch lengths of each Hg subtree, rather than on a specific internal structure. This approach is expected to do well in the presence of a detailed Hg phylogeny, even with relatively small sample.

Demerits: The technique is limited by its assumption on the Y-STR mutation process and the symmetric nature of mutation count. The probability of change in the repeat count for each STR has been considered to be fixed, independent of the nature of change and current repeat count. This is unrealistic on the account of lacking stationary distribution of repeat counts. Hence, this leads to unrealistic STR lengths of infinity or zero.

F) METHOD VI: HMM QUANTI-SNP

QuantiSNP [17] approaches with probabilistic quantification of state classifications and improvement of the accuracy of segmental aneuploidy identification and mapping. Objective Bayes measure is employed to fix the parameters involved in calibrating the model to fixed false positive error rate using re-sampling and feedback.

The *priori* probability that some genetic event (state change) occurs between adjacent SNP loci a distance d apart is modelled as:

$$\rho = \frac{1}{2} \left[1 - \exp\left(-\frac{d}{2L}\right) \right] \quad 1$$

where L is a characteristic length which could either be inferred directly from the data, or adjusted to calibrate the model to a given false positive rate.

Hidden state, z	Copy number, $c(z)$	Number of genotypes, $K(z)$	Description
1	0	0	Full deletion
2	1	1	Single copy deletion
3	2	3	Normal (heterozygote)
4	2	2	Normal (homozygote)
5	3	4	Single copy duplication
6	4	5	Double copy duplication

Table 3

[Hidden States, Copy Numbers and Biological Interpretations]

The transition matrix of hidden states between adjacent SNPs i, j is given by:

$$p(z_{i+1} = j | z_i = i) = \begin{cases} \rho/(N_s - 1), & i \neq j \\ 1 - \rho, & i = j, j \neq \{3, 4\} \\ h(1 - \rho), & i = j, j = 3 \end{cases}$$


where h is the rate of heterozygosity which we set as $1/3$ and N_s is the number of hidden states. The emission probabilities are modelled from Gaussian basics integrated into uniform distributions:

$$p(r|z, \theta) = \pi_r/2R_{\max} + (1 - \pi_r)G(r; \mu_{r,z}, s_{r,z})$$

Dirichlet prior is used for B allele frequency mixture weights. The parameters are trained for minimum false positive error using Viterbi algorithm and Expectation Maximization (EM) objective learning. The Viterbi algorithm applies Bayes Factor BF for each aberration in SNP region I to j with copy number k as shown:

$$BF = \frac{p(\mathbf{r}, \mathbf{b} | \mathbf{z}_{i:j} = k)}{\sum_{\mathbf{z}_{i:j} \neq k} p(\mathbf{r}, \mathbf{b} | \mathbf{z}_{i:j})}$$

Merits: This objective learning algorithm has the following advantages:

- Bayes Factor is the probabilistic measure of the presence if CNV in a SNP region.
- This is one of the highest statistically accurate algorithms with very low false positives owing to extensive training.
- This is among the first applications of OB-HMM to high-throughput genomic datasets and can be easily extended to shared CNV 

Demerits: Though this algorithm has high accuracy rate and is very reliable, the main problem is the extensive time involved in training the parameters using objective learning. The algorithm has to be trained until the parameters to the convergence value.

G) METHOD VII: CNV Finder

The CNV Finder [18] is a variance based automatic CNV detector. This is based on two hypothesis on the ratio variances of the array experiments. The first assumption is that most of the observations are normally distributed around a \log_2 ratio of zero, which is the normal diploid copy number in test and control genomes. The second one is based on the variation in CNV ratio values which fall outside the central distribution. The final calling algorithm allows restricted extension of called regions with ratios greater than thrice standard deviation and permits the incorporation of single, non-consecutive uncalled clones within the region

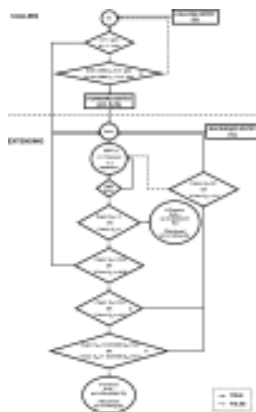


Figure 7 [CNV Finder Algorithm]

Merits: This algorithm has the following merits:

- CNV Finder has very low false positive error rate, lesser than 5% in most cases.
- Compared to SW, it has a very high sensitivity and accuracy rate

Demerits: Due to varying repeat content, sequence homologies, and experimental variation, some clones under-respond to a specific copy number change and may fail to be called in higher SD experiments, thus fragmenting the CNV despite the final iteration effects.

H) METHOD VIII: PennCNV

PennCNV [19] is an integrated HMM for CNV detection using Illumina high-density SNP genotyping with integration from multiple source ranging from total signal intensity and allelic intensity ratio at each SNP markers, intermarker distances, frequency of SNP and distance between adjacent SNPs.

PennCNV incorporates several components together into a hidden Markov model (HMM), including the LRR (log R ratio), the BAF (B allele frequency), the distance between neighboring SNPs, and the population frequency of the B allele as demonstrated in the flowchart of figure [8]. The distance between neighboring SNPs determines the probability of having a copy number state change between them. Also, the family genetic information is integrated to yield better sensitivity with respect to CNV identifications.

Most of the mathematical formulas to calculate emission probabilities of LRR and BAF and posterior probability coincide with that of the HMM explained previously except for the difference in parameter training by Baum-Welch algorithm [20] instead of Viterbi.

Merits: The technique employs Illumina Infinium platform which has the following merits benefitting high resolution CNV detection:

- The assay combines specific hybridization of genomic data to array of probes, thus resulting in higher SNR.

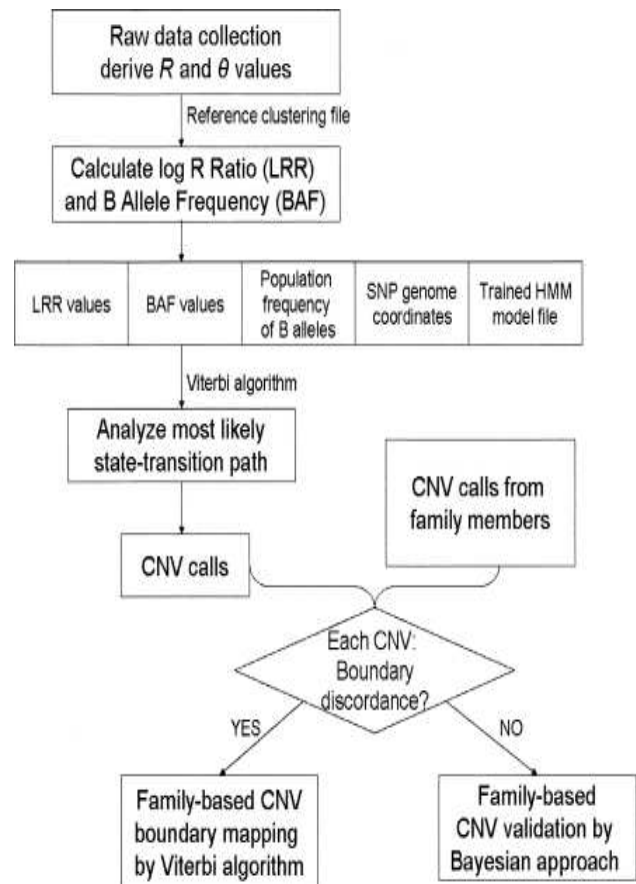


Figure 8
[Flowchart depicting algorithm of PennCNV]

- The assay does not need amplification for PCR and hence, differential amplification of genomic region result in lesser discrepancies.
- The algorithm has added advantage of distinguishing fine scaled CNVs of median size approximately 12kb, smaller than other experiments by an order of magnitude and thus achieves higher resolution.

Demerits: This algorithm has the regular defects associated with that of HMM along with the added ones as listed below:

- The HMM portion requires lot of investment to train the parameters.
- Accuracy is decreased due to the usage of linkage disequilibrium information which results in loss of small CNVs spaced far apart; and also due to the decreased accuracy of clustering file
- The paper employs database that lacks coverage on heterochromatin areas and centromeric regions.
- Due to algorithm used and Hardy-Weinberg equilibrium principle employed in calculations, SNPs with common CNVs are under-represented

VII. PARAMETERS OF EVALUATION AND PERFORMANCES

The popular CNV detection and frequency estimation algorithms have been discussed in detail in the preceding section. This section is dedicated to wrapping up the evaluation parameters involved in the choice of the right algorithm for a particular application. Furthermore, the section also compares the relative performances of the algorithms discussed earlier with statistics emerging from different publications, generalized onto a common platform.

This paper is dedicated to comparative analyses of different available algorithms. This is impossible without characterizing the standards on the basis of which the algorithms should be compared. Below are listed a few parameters which are to be kept in mind while choosing the algorithm for any application.

- ❖ Optimal parameter setting and assumption on the inputs. This is an important parameter as while training parameters and applying theory, some specific assumptions are made about the distribution of CNV, their inter-distances. These might be unrealistic leading to poor performance of algorithms in real experiments. Based on the hypothesis that calls to same genomic region represent false positives in most cases, hence a new parameter normalized singleton ratio (NSR) is defined.

$$NSR = \frac{p_u}{\mu_{cs}}$$

Where p_u is the proportion of unique CNV found in only one sample to μ_{cs} , the average number of CNV SNPs called per sample. The lower the value, the better the technique.

- ❖ Sensitivity, specificity, false positive rates and ROC curve: Both false negatives and false positives are destructive in the results as they might lead to false associations of genetic diseases with specific CNVs. The residual for ROC curve is calculated using the following formula, where larger values are considered optimal [16]:

$$\frac{\sqrt{2}(\text{sensitivity} - (1 - \text{specificity}))}{2}$$

- ❖ Accuracy and prediction rate in terms of correct prediction of boundaries across different CNV sizes: it is important for the algorithm to detect CNVs across different chromosomal regions independent of heterozygosity and distribution in physical domain.
- ❖ Resolution rate: again the CNVs detected should be independent of the sizing as scientists may want to isolate different CNVs for different diseases and distinguish among multi gene spanning CNV
- ❖ Resources Investment: Time and cost is an effective factor as the algorithm must be practically implementable. Ones which take lot of time to train parameters might have a negative bias.

The weighted combination of these parameters helps in deciding the suitability of a particular technique to the given situation. Keeping these parameters in mind and choosing some of the representatives of the techniques discussed above, the remaining part of the section presents their comparison based on acquired and generalized statistics. From the statistics acquired, Quanti SNP seems to be among the best performers in most parameters and a safe choice for estimating CNV locus and frequency. Graphs are shown below:

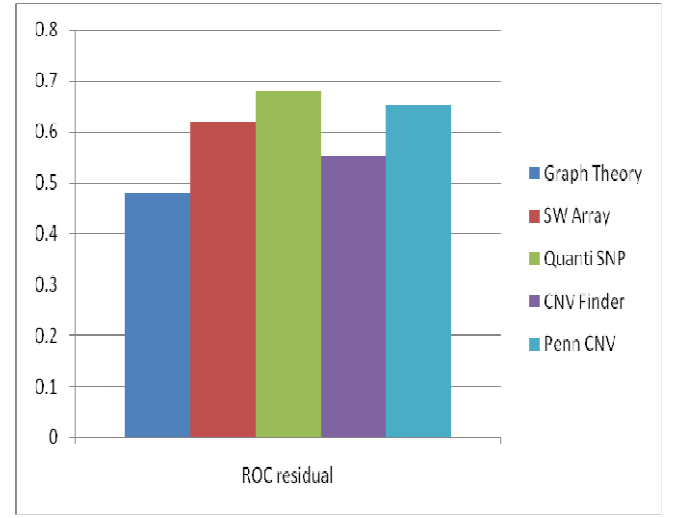


Figure 9

Quanti SNP has the best ROC statistics indicating very low false positives, on the other hand Graph theory exhibits poor ROC residue on account of false predictions.

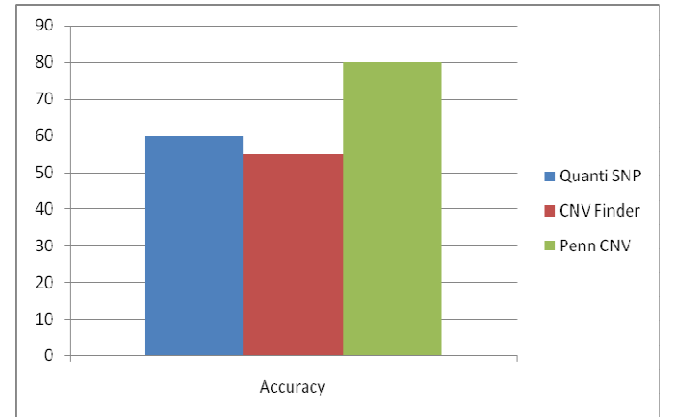


Figure 10

The exact statistics was not available for all techniques on the same scale. But, among the given data, these three represent the more accurate version. While, Bayes based algorithm also leads to higher accuracy, the Graph theory compromise on accuracy for more simplistic procedure.

VIII. FUTURE WORK

As discussed in the earlier section, many other interesting algorithms exist that can be compared on improved parameters by equalizing the sample and generating uniform experiments. Also, the role of these algorithms in predicting phylogenetic relationships and probability of clinical diseases is an interesting application for further research.

ACKNOWLEDGMENT

Biology has always been my favourite course and I aspire to pursue research in an area that integrates electronics with biology. I am extremely thankful to Dr. Douglas Brutlag for offering such a wonderful course and giving me this opportunity to learn about computational aspects of Biology. I express my sincere gratitude to Daniel Davison for his guidance and support in strengthening the understanding of the concepts. This acknowledgment would be incomplete without thanking my family and friends for the moral support that they've always bestowed on me.

REFERENCES

- [1] Raoul-Sam Daruwala, Archisman Rudra, Harry Ostrer, Robert Lucito, Michael Wigler, and Bud Mishra, A versatile statistical analysis algorithm to detect genome copy number variation, *Proc Natl Acad Sci U S A*. Nov 2004 [PMCID: PMC528962].
- [2] Jie Chen, Ayten Yiğiter, Yu-Ping Wang, and Hong-Wen Deng, A Bayesian Analysis for Identifying DNA Copy Number Variations Using a Compound Poisson Process, *EURASIP J Bioinform Syst Biol*. 2010 [PMCID: PMC2952792]
- [3] Iuliana Ionita-Laza, Nan M. Laird,¹ Benjamin A. Raby, Scott T. Weiss, and Christoph Lange, *On the Frequency of Copy Number Variants*, Oxford University Press, 2008 [PMCID: PMC2562008]
- [4] C.C.Mundt, Probability of Mutation to Multiple Virulence And Durability of Resistance Gene Pyramids, *The American Phytopathological Society*, vol. 80, 1990
[http://www.pnpgg.org/pp590_790/phytopathology-letter.pdf]
- [5] J.B.S.Haldane, The Rate of Spontaneous Mutation of a Human Gene, *Indian Academy of Sciences*, vol. XXXI, Oct 1935
[<http://www.springerlink.com/content/817340k38w9gk3/>]
- [6] <http://hstalks.com/dl/handouts/HST81/2328.pdf>
- [7] <http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/M/Mutations.html>
- [8] <http://webcache.googleusercontent.com/search?q=cache:aZ3xaEMl8IJ:www.newton.ac.uk/programmes/CGR/seminars/071310454.ppt+Copy+Number+Variant+as+Poisson+wrong&cd=5&hl=en&ct=clnk&gl=us>
- [9] Thomas S. Price, Regina Regan, Richard Mott, Åsa Hedman, Ben Honey, Rachael J. Daniels, Lee Smith, Andy Greenfield, Ana Tiganescu, Veronica Buckle, Nicki Ventress, Helena Ayyub, Anita Salhan, Susana Pedraza-Diaz, John Broxholme, Jiannis Ragoussis, Douglas R. Higgs, Jonathan Flint, and Samantha J. L. Knight, SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data, *Nucleic Acids Res*. 2005. [PMCID: PMC1151590]
- [10] Ben-Yaacov E, Eldar YC, A fast and flexible method for the segmentation of aCGH data, *Bioinformatics*. 2008 Aug; [PMID: 18689815]
- [11] Graeme Hodgson, Jeffrey H. Hager, Stas Volik, Sujatmi Hariono, Meredith Wernick, Dan Moore, Donna G. Albertson, Daniel Pinkel, Colin Collins, Douglas Hanahan³ & Joe W. Gray, Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas, *Nature Genetics*, Nov 2001
[<http://www.nature.com/ng/journal/v29/n4/full/ng771.html>]
- [12] Xiao Lin-Yin, Detecting Copy Number Variations from Array CGH Data Based On A Conditional Random Field Model, *JBCB*, vol. 8, Oct 2009
[<http://www.worldscinet.com/jbcb/08/0802/S021972001000480X.html>]
- [13] Jie Chan, Yu-ping Wang, A Statistical Change Point Model Approach for the Detection of DNA Copy Number Variations in Array CGH Data, *IEEE-ACM*, vol. 6, Oct 2009
[<http://www.computer.org/portal/web/csd/doi/10.1109/TCBB.2008.129>]
- [14] Jonathan R. Pollack, Therese Sørlie, Charles M. Perou, Christian A. Rees, Stefanie S. Jeffrey, Per E. Lønning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale, and Patrick O. Brown, Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors, *Proc Natl Acad Sci U S A*. 2002 October [PMCID: PMC130569]
- [15] Osnat Ravid-Amir and Saharon Rosset, Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats, *Bioinformatics*. 2010 Sep. [PMCID: PMC2935444]
- [16] Andrew E. Dellinger,¹ Seang-Mei Saw,² Liang K. Goh,³ Mark Seielstad,⁴ Terri L. Young,^{1,3,5} and Yi-Ju Li, Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays, *Nucleic Acids Res*. 2010 May. [PMCID: PMC2875020]
- [17] Stefano Colella, Christopher Yau, Jennifer M. Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S. Bassett, Anneke Seller, Christopher C. Holmes, and Jiannis Ragoussis, QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data, *Nucleic Acids Res*. 2007 Mar, [PMCID: PMC1874617]
- [18] Heike Fiegler, Richard Redon, Dan Andrews, Carol Scott, Robert Andrews, Carol Carder, Richard Clark, Oliver Dovey, Peter Ellis, Lars Feuk, Lisa French, Paul Hunt, Dimitrios Kalaitzopoulos, James Larkin, Lyndal Montgomery, George H. Perry, Bob W. Plumb, Keith Porter, Rachel E. Rigby, Diane Rigler, Armand Valsesia, Cordelia Langford, Sean J. Humphray, Stephen W. Scherer, Charles Lee, Matthew E. Hurles, and Nigel P. Carter, Accurate and reliable high-throughput detection of copy number variation in the human genome, *Genome Res*. 2006 Dec. [PMCID: PMC1665640]
- [19] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F.A. Grant, Hakon Hakonarson, and Maja Bucan, PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data, *Genome Res*. 2007 Nov, [PMCID: PMC2045149]
- [20] Baum L.E., Petrie T., Soules G., Weiss N., Petrie T., Soules G., Weiss N., Soules G., Weiss N., Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math. Statist*. 1970;41:164-171.
- [21] Dmitry Pushkarev, Norma F Neff & Stephen R Quake, Single-molecule sequencing of an individual human genome, *Nature Biotech* 27, 2009 Aug. [<http://www.nature.com/nbt/journal/v27/n9/full/nbt.1561.html>]
- [22] Haiying Huai and R. C. Woodruff, With the Correct Concept of Mutation Rate, Cluster Mutations Can Explain the Over dispersed Molecular Clock, *Genetics*, Vol. 149, 467-469, May 1998
- [23] Stephen W Scherer, Charles Lee, Ewan Birney, David M Altshuler, Evan E Eichler, Nigel P Carter, Matthew E Hurles, and Lars Feuk, Challenges and standards in integrating surveys of structural variation, *Nat Genet*. 2007 July, [PMCID: PMC2698291]
- [24] Heather C. Mefford and Evan E. Eichler, Duplication Hotspots, Rare Genomic Disorders and Common Disease, *Curr Opin Genet Dev*, 2009 May. [PMCID: PMC2746670]
- [25] F. Speleman, C. Kumps, K. Buysse, B. Poppe, B. Menten, K. De Preter, Copy number alterations and copy number variation in cancer: close encounters of the bad kind, *Cytogenet Genome Res* 2008. [PubMed ID 19287153]
- [26] Ken-Shiung Chen, Prasad Manian, Thearith Koeuth, Lorraine Potocki, Qi Zhao, A. Craig Chinault, Cheng Chi Lee & James R. Lupski, Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome, *Nature Genetics* 17, 154 - 163 (1997)
[<http://www.nature.com/ng/journal/v17/n2/abs/ng1097-154.html>]
- [27] David A Koolen, Lisenka E L M Viissers, Rolph Pfundt, Nicole de Leeuw, Samantha JL Knight, Regina Regan, R Frank Kooy, Edwin

- Reyniers, Corrado Romano, Marco Fichera, Albert Schinzel, Alessandra Baumer, Britt-Marie Anderlid, Jacqueline Schoumans, Nine V Knoers, Ad Geurts van Kessel, Erik A Sistermans, Joris A Veltman, Han G Brunner & Bert B A de Vries, A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism, *Nature Genetics* - **38**, 999 - 1001 (2006) [<http://www.nature.com/ng/journal/v38/n9/full/ng1853.html>]
- [28] James R. Lupski, Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits, Volume 14, Issue 10, 1 October 1998
- [29] <http://www.detectingdesign.com/dnamutationrates.html>
- [30] George H. Perry, Fengtang Yang, Tomas Marques-Bonet, Carly Murphy, Tomas Fitzgerald, Arthur S. Lee, Courtney Hyland, Anne C. Stone, Matthew E. Hurles, Chris Tyler-Smith, Evan E. Eichler, Nigel P. Carter, Charles Lee, and Richard Redon, Copy number variation and evolution in humans and chimpanzees, *Genome Res.* 2008 Nov, [PMCID: PMC2577862]
- [31] Gregory M Cooper, Deborah A Nickerson & Evan E Eichler, Mutational and selective effects on copy-number variants in the human genome, *Nature Genetics* 39, S22 - S29 (2007)