On a novel coalescent model for genome-wide evolution of Copy Number Variations

Antonina Mitrofanova* and Bud Mishra

Department of Computer Science,
Courant Institute of Mathematical Sciences,
New York University,
New York, NY, 10003, USA
Fax: (212)998-3484
E-mail: antonina@cs.nyu.edu
E-mail: mishra@nyu.edu
*Corresponding author

Abstract: Since we are limited in our knowledge of human demographic history and variations of recombination and mutation rates, large-scale computer simulation is a necessary tool in genetics. Here, we propose and computationally simulate a model of evolution for unique and segmentally duplicated regions of human genome. Since such segmentally duplicated regions show a complex behaviour of copy number changes, our model is hoped to lead to a better understanding of the evolutionary developments of CNVs, algorithms for associations studies with CNV markers, and finally, for characterising parameters for stochastic diffusion models, describing asymptotic behaviour of evolutionary processes.

Keywords: SDs; segmental duplications; copy number polymorphism; coalescent process.

Reference to this paper should be made as follows: Mitrofanova, A. and Mishra, B. (xxxx) 'On a novel coalescent model for genome-wide evolution of Copy Number Variations', *Int. J. Data Mining and Bioinformatics*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Antonina Mitrofanova is pursuing her PhD in Computer Science at New York University, Courant Institute of Mathematical Sciences, under Professor Mishra's supervision. Her research interests include evolutionary models in population genetics, stochastic processes, copy number variations, probabilistic models for inferring protein classification, graph algorithms, and model checking on graphical models. Before coming to NYC, she spent four years studying medicine at the National Medical University in Kiev, Ukraine. She is a recipient of NYU's Sandra Bleistein Award, CRA-W distributed mentor project award, Stewart M. Monchik Memorial Scholarship, Honorable Mention in NSF Fellowship award and Jack Wolfe Award.

Bhubaneswar Bud Mishra is a Professor of Computer Science and Mathematics at NYU's Courant Institute, Professor of human genetics at Mt. Sinai School of Medicine, and a Professor of cell biology at NYU School of Medicine. He has a Degree in Physics from Utkal University, in Electronics and Communication Engineering from IIT, Kharagpur, and MS and PhD Degrees in Computer Science from Carnegie-Mellon University. He is a fellow of both IEEE and ACM and also, a NYSTAR Distinguished Professor (2001).

1 Introduction

Genetic variations in the human genome take many different forms, ranging from many large chromosomal rearrangements to numerous isolated discrepancies in just single nucleotides. This paper focuses on population genetics of copy number polymorphisms, a recently discovered genetic variation in human genome, which varies from one kb to larger in size and shows copy number differences from individual to individual. Of course, in a diploid normal (but not necessarily the one with dominant allelic form) genome, any such segment is expected to have exactly two copies; but across the population, it may exhibit altered copy number: namely, a reduced copy number, e.g., zero or one, suggesting homozygous or hemizygous deletions, respectively, or an increased copy number, e.g., three or more, suggesting amplifications. Such variations, also called CNVs, remain poorly characterised in terms of their structural properties and evolutionary history. Many CNVs are thought to correspond to normal variations among genomes of individuals with little or no effect on their phenotype. However, some of the copy number changes have been shown to associate with genetic diseases and predisposition to various conditions, such as autism susceptibility, cervical cancer and other tumours; see Makni et al. (2000), Cahill et al. (1999), Duensing and Munger (2003) and Lamb et al. (2005). Thus, it is of immense importance to rigorously study the nature of CNVs, precisely quantify their impact on phenotype of the carrier, and investigate computationally what evolutionary mechanisms may be responsible for producing and maintaining such variations in human genome.

So far, from the data collected by genomic arrays, it seems that CNVs are not uniformly distributed in the human genomes. For example, many studies have observed the enrichment of CNVs in the regions of segmental duplications; see Sharp et al. (2005) and Locke et al. (2006). Additionally, Redon et al. (2006) noticed that various CNVs, specifically multi-allelic, complex and combined deletion-amplification CNVs, are markedly enriched for Segmental Duplications (SDs).

At the same time, our analysis, reported here, has gone further and observed that CNVs in segmentally duplicated regions dramatically differ from those located in unique regions of the genome. For instance, CNVs show unusually different structural patterns depending on whether they occur in unique or segmentally duplicated regions, and do so in widely separated human populations. For instance, we have observed the following:

In general, copy number changes in a genomic region can be roughly characterised as

1 amplifications

2 deletions

3 mixed (some individuals having amplifications and other individuals having deletions).

Our statistical analysis suggests that unique and segmentally duplicated regions show different distributions in terms of these characteristics. In particular, unique regions of the genome can have either 1) or 2), but not 3) (either amplification or deletion, but not both). In contrast, segmentally duplicated regions experience all of the patterns: 1), 2) and 3). In other words, all cases of 3) (mixed amplifications and deletions) are primarily located in segmentally duplicated regions.

Motivated by the observations above, we hypothesised that perhaps a different nature of evolution of CNVs dominates in segmentally duplicated regions, but not seen in unique regions. We may thus expect to be able to discern distinct mechanisms of evolutionary development of CNVs operating differently for unique and segmentally duplicated regions of genome by designing coalescent processes for both of them appropriately.

It is a necessary and crucial computational task in population genetics to produce a faithful simulation for the models of evolutionary development and validate them against real population data. These models will not only enable a scientist to observe and analyse the assumed behaviour, but also help establish a strong link between current and next steps in evolutionary population research as well as in biomedical applications of genetics.

In preparation for our models, we first briefly describe the previous relevant research examining the connection of CNVs and regions of SDs.

2 Regions of Segmental Duplications vs. unique regions: motivation for a new coalescent model

Segmental Duplications (SDs) are 1–400 kb long highly homologous blocks of DNA that occur at more than one site within the genome. Regions of SD in the genome have been known as regions prone to chromosomal instability, recombination and recurrent chromosomal rearrangements associated with certain genomic diseases. These regions have attracted interest of many scientists as their connections to CNVs have become apparent; see the discussions in Sharp et al. (2005), Locke et al. (2006) and Redon et al. (2006).

For instance, Sharp et al. (2005) and Locke et al. (2006), noticed a significant enrichment of CNVs in the neighbourhood of recombination hot-spots, as compared to any arbitrary control region in the genome. The authors thus concluded that SDs might be major mediators of large-scale variation in the human genome. In particular, the study of Sharp et al. (2005) suggested that certain regions of the genome are predisposed to rearrangements, and most specifically to CNVs. The best examples of such regions are seen in the genomic regions containing a SD.

All of these preceding observations encouraged us to study CNVs, and how their behaviours vary in segmentally duplicated as compared to unique genome regions. We analysed HapMap data of unrelated individuals from China (46 people) and Japan (45 people) as well as Trios from Yoruba (89 people, 29 trios) and Ceph (Utah of European origin, 90 people, 30 trios). Our analysis here is based on existing Bacterial Artificial Chromosome (BAC) array data, which were kindly provided to us by Drs. Evan Eichler and Andy Sharp (duplicate clones were ignored in our study).

To distinguish CNVs from normal copy values we have used a criterion similar to the one used by Sharp et al. (2005) and Locke et al. (2006). In particular, to account for asymmetry in some hybridisation data, we used the following statistical method to identify variants in an asymmetric distribution. The total distribution of log_2 ratios was divided into two groups by their averages. The standard deviation from the average was then determined for above average and below average groups *separately*. The data for each group was mirrored to simulate a symmetric distribution. The variant threshold for gains was calculated by adding 2 SD of the above-average group to the mean; and similarly, for losses.

2.1 Unique regions

We define unique regions as those having no inter- or intra- SDs. It was observed that whenever unique regions experience CNVs, it is either a deletion or an amplification (homogeneous changes) event, and almost never both (heterogeneous changes). However, there are negligibly few exceptional cases, where regions show deletions and amplifications at the same time (shown as mixed in the Table 1). We examined each such case individually, as described in a footnote[1] below. We also summarise all observed CNV changes for various populations in a table (see Table 1).

Table 1 CNV changes in non-segmentally duplicated (unique) regions of the human genome. The regions are represented by clones. We have used the * mark to indicate the region on chromosome X which is excluded from the subsequent analysis

	Yoruba	*Japanese*	*Chinese*	*Ceph*
No polymorphism	810	817	817	799
Amplifications only	43	43	46	55
Deletions only	46	37	36	44
Mixed	1*	$3 = 2 + 1$*	1*	$2 = 1 + 1$*

All of those cases, described above, seem to be exceptions from the general CNV pattern observed in unique regions of the genome. Thus, it led us to conclude that unique regions show a clear picture of homogeneous polymorphism (either deletions or amplifications, but not both at the same chromosomal location) with rather few exceptions discussed above. We use these observations to propose a model of evolution for CNVs for unique regions of the genome in Section 3. Note that with significantly more and higher resolution data, it will be possible to strengthen the underlying assumptions further, and add more details to the modelled mechanisms.

2.2 Segmental Duplication regions

Our observations indicate that, in contrast to unique genomic regions with a simple form of CNVs, in regions containing a SD, the CNV changes often have heterogeneous character (Table 2). In other words, in the neighbourhood of regions containing SDs, we found that both amplifications and deletions co-occur significantly frequently, even when a single population is examined. Moreover, we observed enhancement of such heterogeneous, mixed polymorphic changes in segmentally duplicated regions in all of the populations (see Table 2 for details).

Table 2 Polymorphic changes in segmentally duplicated regions (clones)

	Yoruba	*Japanese*	*Chinese*	*Ceph*
No polymorphism	786	794	785	741
Amplifications only	124	135	141	129
Deletions only	101	86	101	141
Mixed	43	40	27	44

We used a Fisher exact test to show that ‘mixed’ CNV changes occur significantly more frequently in SD regions than in Unique regions. The exact p-values are: for Yoruba $p = 7.6 \times 10^{-13}$, Japanese $p = 5.7 \times 10^{-10}$, Chinese $p = 8.1 \times 10^{-8}$, and Ceph $p = 1.6 \times 10^{-12}$.

Such differences in CNV behaviour suggest a different mechanism of evolution for CNVs in segmentally duplicated regions. We hypothesise this mechanism below and validate it through large-scale computational simulation.

Additionally, we examined CNVs in the regions of SD, comparing their distributions in inter- vs. intra-chromosomal duplications. In the regions where the CNVs are heterogeneous, we found them in the presence of both inter- and intra-segmental duplications (on the same clone). However, the inter-segmental duplications usually represent regions with mixed CNVs with a significance dominance of either just amplification or just deletion events (for example, 12 amplifications and 2 deletions). At the same time, the intra-segmental duplication showed a slight dominance toward equally-mixed CNVs (with almost equal number of amplifications and deletions in the same clone). As a result, we are likely to ascribe heterogeneous CNV changes to *intra-segmental duplications* (but however, do not exclude the role of inter-chromosomal duplications in some cases of the mixed CNVs). Thus, our model of evolution for segmentally-duplicated regions (described in Section 3) assumes intra-segmental duplications.

3 Coalescent model

Motivated by the above observations, we created separate evolutionary models for unique and segmentally duplicated regions, suggesting different mechanisms for the development of CNVs in these regions.

The starting point of our development is a Moran model in a population of constant size as described in Moran (1962). At each time step, a randomly selected representative leaves two offspring, while it and another randomly selected representative die, thus keeping the population size conserved. In Moran model, as opposed to Wright-Fisher model proposed in Fisher (1922, 1930) and Wright (1931, 1945), populations overlap thus presenting a more realistic conditions for a faithful simulation.

We assume that the population in our simulation consists of **N** haploids. We fix some region of the genome, for which the sequence is assumed known. It is modelled either as a unique region or as a region of SD, as discussed in detail below.

We assume further that this genomic region possesses an allele (ancestral short sequence segment that would eventually give rise to a multi-allelic CNV), whose copy number can be estimated. We wish to trace the evolution of this region in the sample of **n** haploids. In particular, we build a coalescent tree tracing the history of the sample back in time, up to the Most Recent Common Ancestor (MRCA).

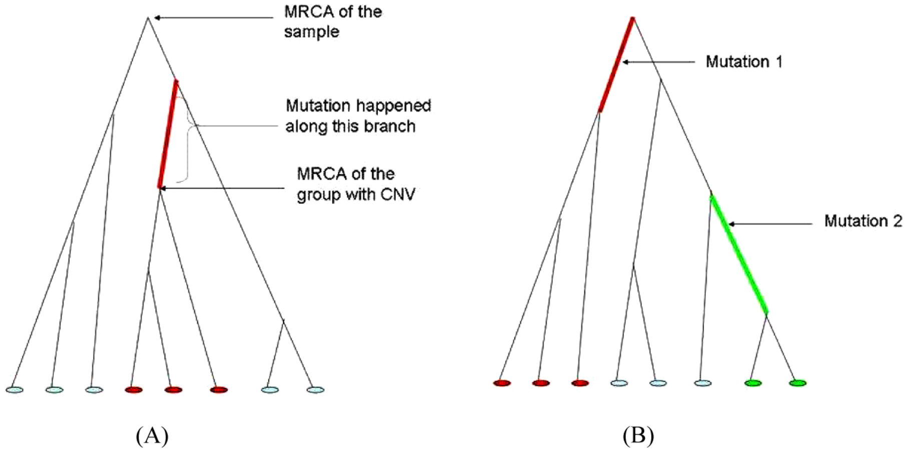
At a certain instant during the evolution, the ancestral allelic form can gain one more copy (get amplified) or lose one copy (get deleted). We generalise this process and call it generically a 'mutation', without specifying the mechanism responsible for the ancestral segment to get amplified or deleted.

3.1 Evolution of unique regions

We fix some unique region of the genome, which has no inter- or intra-chromosomal duplications. Notice that since in the unique regions of the genome the copy number changes can be assumed to have a homogeneous character (either toward amplification or toward deletion), the mutation event for this region is either amplification or deletion, but never both. In other words, in the coalescent model of a specific region there will be only one type of mutational changes (say amplification).

The model assumes that certain specific 'mutation' had occurred once in the past and then are 'propagated' down the descendant subtree. Thus, for the group of haploids possessing a specific mutation, the mutation event happened in their MRCA. For examples, see Figure 1(A). We assume that there are no other additional mutations in that descendant (red) subtree. Said another way, if a specific unique region has experienced mutation once, the probability that it will suffer the same fate again is close to zero.

Figure 1 Coalescent tree: (A) mutation happened once along the red branch and then was inherited on the descendant subtree and (B) after several mutations, each along the red and green branches: The mutations are propagated on the descendant subtrees. We assume that the descendant subtrees acquire no new mutations (see online version for colours)


MRCA of the sample
Mutation happened along this branch
MRCA of the group with CNV
(A)
Mutation 1
Mutation 2
(B)

However, we do not limit our coalescent process to just a single mutation for the whole sample. In particular, several mutations are allowed to occur along the branches as the evolution progresses, as shown in Figure 1(B). Each of these mutations propagates down its own descendant subtree. This process results in several subgroups in the sample (e.g., red and green) having different MRCAs that initiated their mutations. Again, we assume it unlikely to have an additional mutation within any of these subtrees.

3.2 Evolution of segmentally duplicated regions

The nature of CNVs in the region of SD reveals more complex evolutionary mechanisms at play.

In segmentally duplicated regions, the copy number changes in the same region can be represented by amplifications in some individuals and deletions in others. Thus, the regions of SD allow mutations leading to amplifications and mutations leading to deletions in the same region.

We propose the following two mechanisms occurring in the coalescent tree and leading to both amplifications and deletions in the same segmentally duplicated region.

3.2.1 First mechanism

First of these mechanisms just assumes some non-zero probabilities for mutation-amplification and mutation-deletion in the same region. We support this assumption by our earlier observation that the regions of segmental duplication are associated with high genomic instability. In this case, as shown in the Figure 1(B), the mutation-amplification can occur along the red line and mutation-deletion, along the green line. In this scenario, some representatives in the population possess amplifications while others have deletions in the same region.

3.2.2 Second mechanism

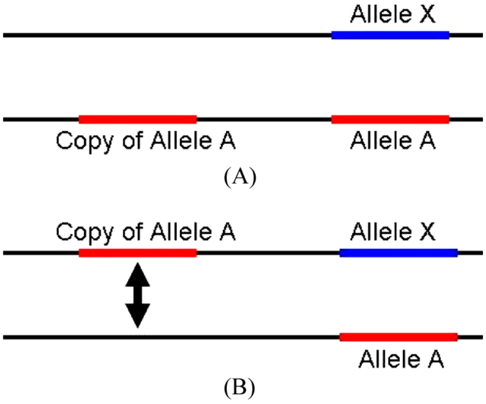
Another mechanism, which we propose, involves the idea that the regions of SD are prone to genomic rearrangements, such as those induced by genetic recombination. We assume the existence of recombination hot-spots in this region, which increase the chances of meiotic recombination.

After the mutation-amplification has occurred in a branch, it propagates down the line of immediate descendants. We may now direct our attention to one of the descendants possessing this amplification, as in Figure 2. In this case, the lower chromosome in Figure 2(A), has two copies of allele A.

The Figure 2(B), shows recombination event occurring during meiosis. Because of recombination, one copy of the allele A gets exchanged with the other (upper) chromosome. In other words, one copy of allele A is ‘lost’ from the lower chromosome.

We may then observe that at this moment the lower chromosome returns back to its normal state, with one copy of the allele A. Such events are considered crucial in our coalescent-tree-based population model. After the lower chromosome has experienced the recombination event, it is left with one (original) copy of allele A. Since the lower chromosome is now in its original unaltered state, in our model, it can now be subjected to amplification or deletion mutations *again*. For example, if the amplified region loses one copy of the allele in the recombination event, then it can lose the original copy of allele A because of the mutation-deletion event. This scenario would correspond to losing allele A in the lower chromosome in Figure 2(B).

Figure 2 Recombination: (A) one chromosome has Allele A amplified (Allele A and its copy); the other chromosome has Allele X ($X \in \{A, B\}$) and (B) after the recombination event: Now, one chromosome has Allele A and another chromosome has Allele A and X (see online version for colours)



We still retain the basic evolutionary properties of the model: namely, if a mutation is imposed on the descendant subtree, an additional mutation on that subtree still remains impossible. However, we modify this assumption when the recombination is involved. After the amplification event, it now becomes possible to have another mutation in the descendant subtree (either amplification or deletion) after the recombination event has brought the region to its native ‘unaltered’ state.

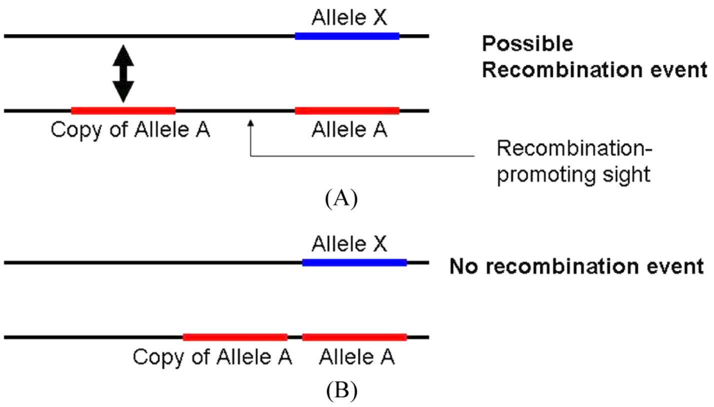
3.2.3 Distance-dependent recombination

We next augment the coalescent tree model appropriately so that the recombination can be treated as dependent on the distance between the two copies of allele A, as in Figure 3. We assume that the regions of SD have multiple sites that promote recombination and that these regions are uniformly distributed along the region. As a result, if the distance between allele A and the copy of allele A is large, then it is more likely that the region between the allele and its copy has one or more recombination-promoting sites, which makes the recombination event likely to happen, and to return the region to its ancestral state. In fact, the relationship between the distance of two copies and the chance of recombination follows exponential distribution.

Therefore, if two copies are far from each other, the linkage between them will most likely be weak as a result of the high chance of recombination. Even further, if two copies are located in different chromosomes, then the distance between these copies is treated as infinity suggesting no linkage between the copies.

Intuitively, as a result of these assumptions, recombination will be seen to promote further amplifications or deletions with a rate that monotonically depends on the distance between copies of an intra-chromosomal duplication. While the model is simple and does not necessarily require conceiving any new and unusual evolutionary mechanism, it already correctly accounts for the differences in CNV distributions in unique vs. segmentally duplicated regions. The model also suggests several falsifiable hypotheses.

Figure 3 Distance-dependent recombination: (A) distance between allele A and its copy is large so that the recombination event is likely and (B) distance between allele A and its copy is small, which makes the recombination event very unlikely (see online version for colours)



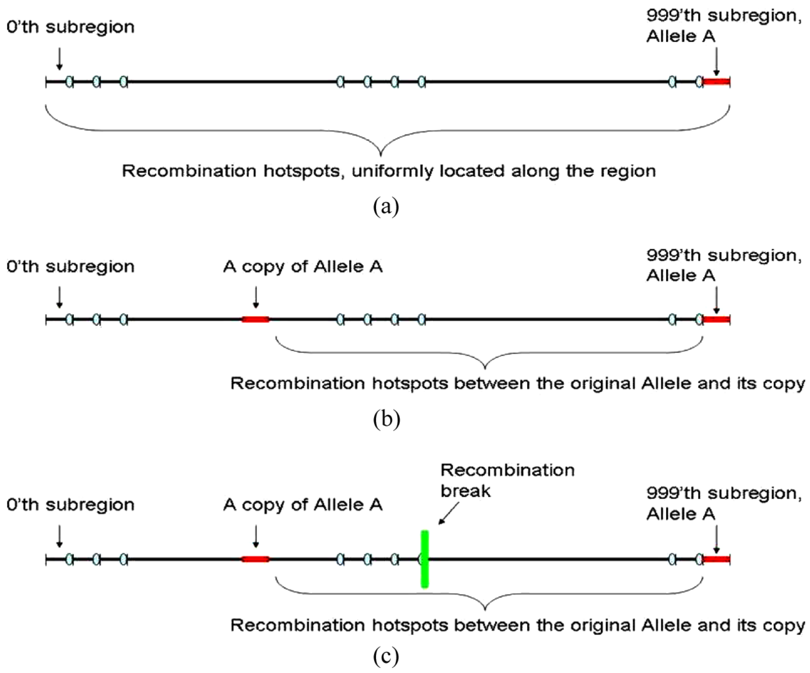
4 Simulation and results

We implemented a software system to simulate a coalescent process in a region prone to SDs (we omit a similar model in unique regions due to its relative simplicity). We simulated a coalescence for a sample of 100 haploids from the population of size 10,000.

We assume that our region contains multiple recombination hot spot motifs. In particular, we 'divide' the regions into 1000 consecutive subregions and assume that hot spot motifs are uniformly located along the region: each of the 1000 subregions contains one hot spot. This assumption can easily be changed and hot spots can be assumed to concentrate in some specific location, say in the centre of the region. In this way, all recombination events (breaks) would happen in the middle of the region. In our current version, the recombination is equally likely to happen near any of recombination hot spots and thus is uniformly distributed along our fixed segmentally duplicated region.

In the simulated (haploid) region, we assume a presence of an allele and fix its position at the right-most distal end: in the 999th subregion. Please see Figure 4(a) for details. There are two variations of allele in place: allele A and allele B. The mutation events (amplification or deletion) are assumed to occur with respect to only one of them (we fix allele A for this purpose) and the other one (allele B) remains unaltered by any mutation process. Saying differently, it is not possible for a haploid to have several copies of allele B (amplification event) as well as allele B deleted (deletion event). However, it is possible for a haploid to possess a copy of allele B and several copies of allele A, which were gained through recombination. On the other hand, allele A is widely exposed to mutations, as described by two proposed mechanisms of evolution in the segmentally duplicated region. it is important to note that only the original allele A, not any of its copy, can be exposed to mutations.

Figure 4 The scheme describes assumptions made in the coalescent simulation software: (a) a region is divided into 1000 subregions; 999th region is assumed to contain an Allele; (b) the hot spot motifs are uniformly distributed along the entire region; whether a copy of Allele A is lost or kept on, depends on the distance from the original Allele: the bigger is the distance, the larger is the change or recombination and (c) possible recombination break shown (see online version for colours)


0'th subregion
999'th subregion, Allele A
Recombination hotspots, uniformly located along the region
(a)
0'th subregion
A copy of Allele A
999'th subregion, Allele A
Recombination hotspots between the original Allele and its copy
(b)
Recombination break
0'th subregion
A copy of Allele A
999'th subregion, Allele A
Recombination hotspots between the original Allele and its copy
(c)

As discussed above, in segmentally duplicated region the mutation can happen in both directions: as amplification and as deletion. The simulation program takes both deletion rate and amplification rate as the input parameters. If mutation is decided to happen, then it is probabilistically assigned to either amplification or deletion. We varied the amplification rate, with the most interesting case being at 0.5 (equal probabilities for amplification and deletion, given that the mutation happened).

If mutation happens along some branch of the coalescent tree, it is propagated as long as the region that contains the mutation (amplification or deletion) is inherited. We assume that as long as the mutated region is inherited, it is *very unlikely* (probability close to zero) for a new mutation to happen in the mutation-possessing haploid.

Consequently, it is very unlikely to have more than one mutation event on the same branch of the coalescent tree. Since mutations occur as a Poisson process along the branches of the tree, with parameter equal to the product of mutation rate and length of the branch, it is possible for more than one mutation to happen on a single branch. However, even if more than one mutation is to happen along the same branch, only one is considered and the rest ignored.

The inheritance of the mutated region, however, can be stopped. For example, it can happen due to a recombination event. Thus it is possible for a mutated region to lose a mutation and come back to its *original* state (with only one of the alleles present in the 999th positions). Additionally, in such a case the crucial assumption is that the haploid acquires ability to experience a new mutation (either amplification or deletion, as determined probabilistically).

In our simulation, 100 haploids are assigned their original alleles (either A or B) in the 999th position. As the coalescent process proceeds, recombination events happen to all haploids: either containing allele A or containing allele B. When the coalescent tree is built (the common ancestor of the sample is found), we run 'mutation process' down the coalescent tree (deciding whether it is amplification or deletion probabilistically). If mutation on Allele A happens to be amplification, the copy of the allele is placed in one of the subregions. Since we consider a distance-dependent recombination, the further apart from the original allele is the copy, the more likely it is to be lost later due to the recombination event. Therefore, the loss of the allele's copy is proportional to the distance to the original allele and follows the laws of exponential distribution: the bigger is the distance between two copies from each other, the greater is the chance of the recombination event in the region separating them.

We considered recombination and mutation rates being constant in the gene reproduction. However, it is possible to change this assumption and re-calculate recombination and mutation rates after each event with diffusion approximation. In our simulation, we varied both rates and observed fluctuating behaviour of the initial sample, as described below.

In Figure 5 we present a summary statistics which describes the parameters observed in the original sample of 100 haploids after simulation, namely, we map the relationship between the recombination rate and the

(A) proportion of haploids which possess mutations

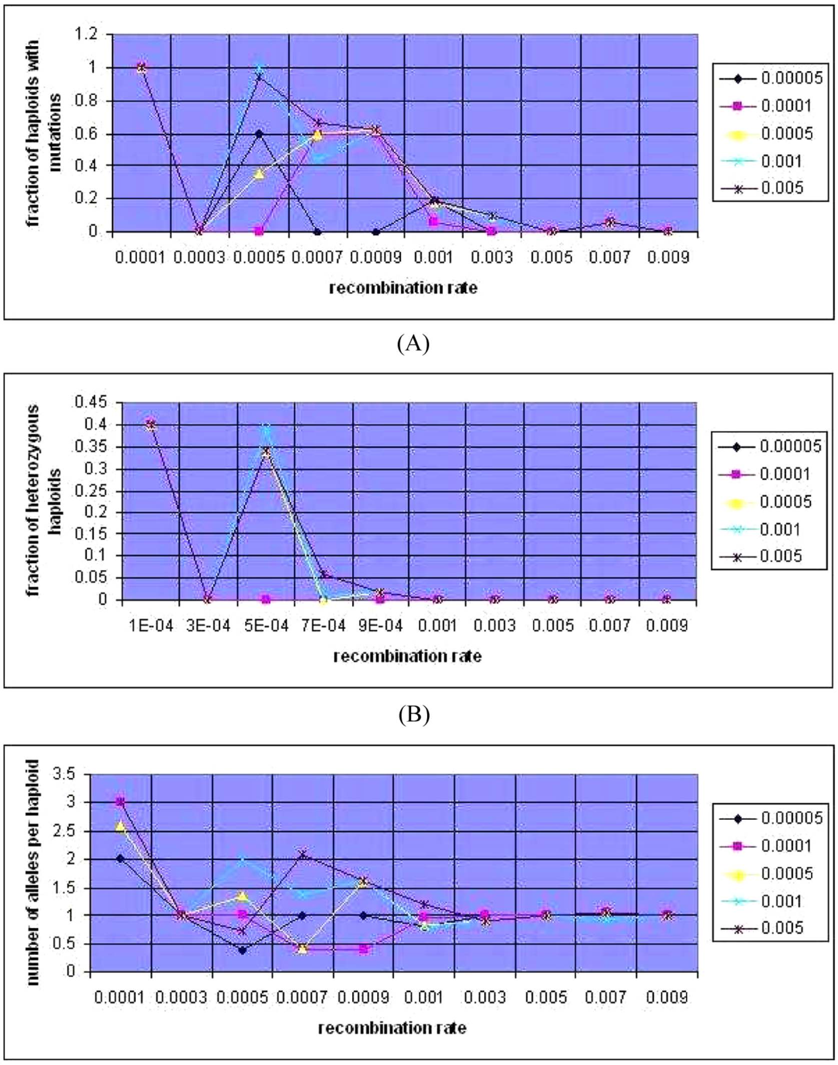
(B) average number of alleles per haploid

(C) heterozygosity (in this case we mean hapoloids which possess both alleles A and B).

We experimented with a variety of cases, changing recombination, mutation and amplification/deletion rates.

Figure 5 compares a summary statistics for simulations of a family of different mutation rates (mutation rates varied from 0.00005 to 0.005, shown as legends). All simulations were performed with the same random seed, thus allowing the analysis to reflect similar history and the variation of summary statistics to reflect the same random mating, while recombination rates varied. In each case, the mutation rate is fixed at one of the specified values while the recombination rate ranges from 0.0001 to 0.01. In this example, we kept the amplification : deletion proportion at 1 : 1 (amplifications and deletions are equally likely, conditioned on the event that the mutation occurred).

Figure 5 Summary statistics for simulations at mutation rates of from 0.00005 to 0.005 per haploid per generation. Relationship between various recombination rates and (A) fraction of haploids with mutations (B) fraction of heterozygous haploids (C) number of alleles per haploid (see online version for colours)



(A)

(B)

(C)

Figure 5(A), shows the increase of the proportion of haploids with mutations when the recombination rates vary from 0.0003 to 0.003. Inside this range, as the recombination rate decreases, the ancestral mutations are more likely to be inherited and in the majority of cases, they are not affected by interference from the recombination events. Consequently, it is not surprising that as a recombination rate decreases (but still remaining non-zero), the number of heterozygous haploid keeps on increasing, as in Figure 5(B). This situation is due to recombination and mutation events near the top of the tree, which then propagate to the descendants with little interference from other subsequent recombinations along the way. Finally, the number of alleles per haploid (Figure 5(C)) fluctuates and heavily depends on the choice of either amplification or deletion causing mutation on the ancestral portion of the tree.

For an example shown in Figure 5, the simulation runs with the recombination rate set to 0.001 and mutation rate to 0.0001 constitute the best fit to the observed real-life data. Note that the mutation rate used here does not refer to a commonly known mutation rate per base, but corresponds to either amplification or deletion of an allele.

Haplotype variations: Because of limitations inherent to the currently available technology, it is not always feasible to find the exact copy number for a certain clone. However, the wide deviation from the average suggests that the copy numbers show a wide range of variations. This effect is consistent with our assumption that it is possible for a haploid to have more than one copy of an allele or a combination of alleles due to amplification, deletion and recombination events.

Our simulation produces various real-life haplotypes. For example, the possible haplotypes of haploids possessing allele A(I) or allele B(II) are shown in the Figure 6.

Allele A can experience either deletion or amplification only when it is in its original state. (It can be at the initial unaltered state or can re-achieve the state after a sequence of recombination events, as can be traced by following the recombination arrows in Figure 6).

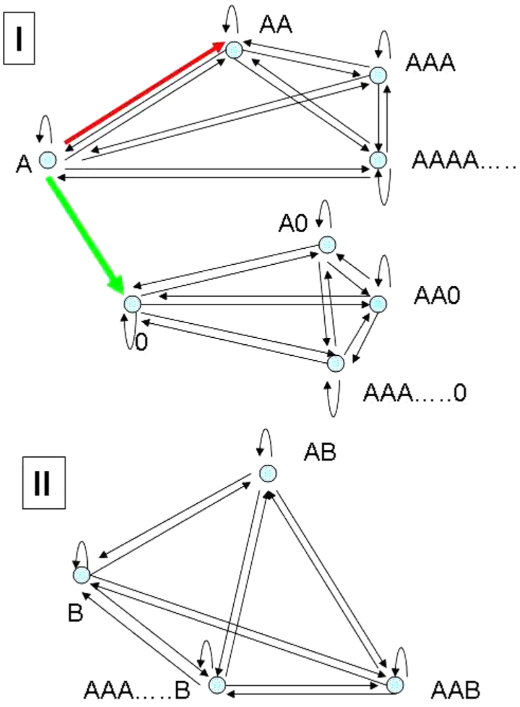
At the same time, Allele B is not influenced by mutation, however the state of a haploid possessing this allele can be altered through the various recombination events (see Figure 6). For example, an allele-B-possessing haploid can gain one (or more) copies of allele A through the recombination with another haploid, which has multiple copies of allele A in place. At the same time, the descendants of this haploid can still lose one or more of these copies of allele A via a different series of recombination events.

4.1 Discussion and conclusions

The models proposed in this work, are powerful enough to explain differences in CNVs in segmentally duplicated vs. unique regions of the genome. We simulated a coalescent process in segmental duplicated region, which involves recombination as well as various mutation events. Even though we simulate a single isolated region of SD, one can observe different behaviours (and thus different regions) through varying mutation, recombination, and amplification/deletion rates. It is easy to adjust our simulation to smaller or bigger number of subregions into which the simulated regions is divided. Additionally, even though the simulated region is assumed to have a uniformly distributed recombination hot spots, it is possible to modify this assumption and place the hot spots in the middle of the region or closer to/further from the allele. Similarly, the position of the allele can be moved to the middle of the region etc.

Note that although in this work we do not discriminate between male and female individuals, in reality, certain traits may be gender specific, and should be accounted for in the simulation. For example, autism is phenotypically manifested primarily in male but not female, as suggested in Lamb et al. (2005). As a result,

Figure 6 Red arrow represents amplification, green arrow – deletion, and black arrow – recombination event. Allele A experiences mutations while allele B remains unaltered. Haploids possessing either allele can gain or lose additional copies of allele A through recombination. (I) shows possible haplotypes of a haploid possessing allele A and (II) shows possible haplotypes of a haploid possessing allele B (see online version for colours)



autistic males usually leave no offspring because of their difficulties in finding mates. On the other hand, females with autism most often exhibit much milder symptoms, usually mate and produce several (often multiple) offspring. In view of these observations, the mating process cannot be assumed random (as we have done here), as it should not allow any descendants for an autistic male in the population.

Furthermore, to account for the inherent asymmetry with which selection affects amplifications and deletions (the former is dominant while the later recessive), the model needs to reflect the fact that a genomic change which provides an individual or sub-population with a selection advantage, are retained over time while the disadvantageous ones vanish (Makni et al., 2000; Cahill et al., 1999; Duensing and Munger, 2003). Therefore, the differential effects of selection pressure for amplification and deletions should be modelled more carefully in the future work.

Additional planned improvements to our model include the following: moving from a segregated population to intra-population mixing, incorporating more realistic conditions of mating and varying the mutational rates to model ‘preferential attachment’. These changes are hoped to further increase the level of interest in these models for population geneticists.

Finally, our simulation together with available population data can be used to perform additional parameter estimation (such as M-statistics, the average extent of linkage disequilibrium etc.) and to connect to various evolutionary scenarios (e.g., population bottlenecks or large-scale migrations); see Pritchard et al. (2003). In this way, our and other similar studies will greatly enhance the understanding of the genome-wide variations, such as variations in copy number, mechanisms underlying their evolutionary history and their possible relation to regions of genomic instability.

References

Cahill, D.P., Kinzler, K.W., Vogelstein, B. and Lengauer, C. (1999) 'Genetic instability and Darwinian selection in tumors', *Trends Cell Biol*, Vol. 9, pp.M57–M60.

Duensing, S. and Munger, K. (2003) 'Mechanisms of genomic instability in human cancer: insights from studies with human papillomavirus oncoproteins', *International Journal of Cancer*, Vol. 109, No. 2, pp.157–162.

Fisher, R.A. (1922) 'On the dominance ratio', *Proc. Roy. Soc. Edinburgh*, Vol. 42, pp.321–341.

Fisher, R.A. (1930) *The Genetical Theory of Natural Selection*, Clarendon Press, Oxford.

Lamb, J.A., Barnby, G., Bonora, E., Sykes, N., Bacchelli, E., Blasi, F., Maestrini, E., Broxholme, J., Tzenova, J., Weeks, D., Bailey, A.J. and Monaco, A.P. (2005) 'Analysis of IMGSAC autism susceptibility loci: evidence for sex limited and parent of origin specific effects', *J. Med. Genet.*, Vol. 42, pp.132–137.

Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M. and Eichler, E.E. (2006) 'Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome', *Am. J. Hum. Genet.*, Vol. 79, No. 2, pp.275–290.

Makni, H., Franco, E.L., Kaiano, J., Villa, L.L., Labrecque, S., Dudley, R., Storey, A. and Matlashewski, G. (2000) 'p53 polymorphism in codon 72 of human papillomavirus-induced cervical cancer: effect of inter-laboratory variation', *International Journal of Cancer*, Vol. 87, No. 4, pp.528–533.

Moran, P.A.P. (1962) *The Statistical Processes of Evolutionary Theory*, Clarendon Press, Oxford.

Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2003) 'Inference of population structure using multilocus genotype data', *Genetics*, Vol. 164, pp.1567–1587.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez, J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W. and Hurles, M.E. (2006) 'Global variation in copy number in the human genome', *Nature*, Vol. 444, No. 23, pp.444–454.

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., Oseroff, V.V., Albertson, D.G., Pinkel, D. and Eichler, E.E. (2005) 'Segmental duplications and copy-number variation in the human genome', *Am. J. Hum. Genet.*, Vol. 77, No. 1, pp.78–88.

Wright, S. (1931) 'Evolution in Mendelian populations', *Genetics*, Vol. 16, pp.97–159.

Wright, S. (1945) 'The differential equation of the distribution of gene frequencies', *Proc. Natl. Acad. Sci.*, Washington, DC, USA, Vol. 31, pp.382–389.

Note

[1] Region of chromosome X, 423, 854–595, 665, shows massive amplifications and deletions within a population. This region is located near the region of SD as reported in Locke et al. (2006). Even more interestingly, this region shows the same heterogeneous patterns in all four populations, suggesting that this polymorphism may be ancient and may have suffered multiple modifications. It also appears likely that the chromosomal instability in this region is common across populations. Thus, we decided to treat this as an exception and exclude from the subsequent analysis of unique regions.

Another anomalous example appears in a region on chromosome X, 151, 546, 640–151, 752, 281. The CNV in this region is highly amplified in Chinese, Japanese, Yoruba, and Ceph populations (30% of people in each population). However, in Ceph, it also experiences just one deletion. Above regions on chromosome X appear relatively close to telomeric regions of the chromosome X, which might be a reason for their instability.

The region on Chromosome Y, 8,238, 490–498, 436, 847, experiences massive deletions in all populations. Additionally, in Japanese population it shows one amplification. Most likely this is due to the high hot-spot content of the region, as reported by Locke et al. (2006). At the same time, Y chromosome can be easily omitted from the analysis due to its high instability (AZF deletions etc).

The last, namely, the fourth, region that shows some deviation from homogeneity of polymorphic changes in unique genome regions is the region on chromosome 8, 2105, 019–022, 279, 635, which also has a high hot-spot content; see Locke et al. (2006). This region has very few polymorphic changes. In particular, it shows two deletions in Yoruba (one of which is due to the somatic mutation) and a single deletion and a single amplification in Japanese population.