STRUCTURAL VARIATION IN THE HUMAN GENOME

by

ANDY WING CHUN PANG

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Department of Molecular Genetics

University of Toronto

STRUCTURAL VARIATION IN THE HUMAN GENOME

ANDY WING CHUN PANG

Doctor of Philosophy

Department of Molecular Genetics

University of Toronto

2013

Abstract

The study of variation found in DNA is fundamental in human genetic studies. Single nucleotide polymorphisms (SNPs) are simple to document because they can be captured in single DNA sequence reads. Larger structural variation including duplications, insertions, deletions, termed as copy number variation (CNV), inversions and translocations are more challenging to discover. Recent studies using microarray and sequencing technologies have demonstrated the prevalence of structural variation in humans. They can disrupt genic and regulatory sequences, be associated with disease, and fuel evolution. Therefore, it is important to identify and characterize both SNPs and structural variants to fully understand their impact.

This thesis presents the analysis of structural variation in the human genome. The primary DNA sample used for my experiments is the DNA of J. Craig Venter, also termed HuRef. It was the first personal human genome sequenced. I combined computational re-analysis of sequence data with microarray-based analysis, and detected 12,178 structural variants covering 40.6 Mb that were not reported in the initial sequencing study. The results indicated that the genomes of two individuals differed 1.3% by CNV, 0.3% by inversion and 0.1% by SNP. Structural variation discovery is dependent on the strategy used. No single approach can readily capture all types of variation, and a combination of strategies is required.

I analyzed the formation mechanisms of all HuRef structural variants. The results showed that the relative proportion of mutational processes changed across size range: the majority of small variants (<1kb) were associated with nonhomologous processes and microsatellite events; median size variants (<10kb) were commonly related to minisatellites and retrotransposons; and large variants were associated with nonallelic homologous recombination.

Eight new breakpoint-resolved HuRef inversions were genotyped in populations to elucidate these understudied variants. I discovered that the structures of inversion could be complex, could create conjoined genes, and their frequencies could exhibit population differentiation.

The data here contributes to our understanding of structural variation in humans. It shows the need to use multiple strategies to identify variants, and it emphasizes the importance to examine the full complement of variation in all biomedical studies.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Stephen Scherer for his confidence in me and, for his support and for exemplifying scientific creativity. I thank my committee members Dr. Michael Brudno and Dr. Peter Ray for their guidance and critiques. I would also like to thank Dr. Lars Feuk and Jeff MacDonald for introducing me to genomics and variation detection, for guidance and encouragement throughout my journey as a PhD student, and for their friendship. I thank Dr. Mohammad Arshad Rafiq for teaching me many laboratory techniques, Dr. John Wei and Dr. Dalila Pinto for their assistance in sequence and microarray analyses. I like to express my gratitude to Dr. Ohsuke Migita for co-leading the inversion genotyping project, for sharing many of his interesting stories, and for his kindness. I'd like to thank Dr. Zhuozhi Wang for his brilliant work in tackling many tough genomic/computational problems, and for teaching me many life wisdoms. I owe my debt to Justin Foong, Lynette Lau and Dr. Ryan Yuen for their hard work in the SOLiD sequencing project; I also appreciate their friendships. I am grateful for students and technicians who assisted me: Amandeep Khera, Michael Gritti, Nan Chen and Dennis Chen. To my peers, Dr. Layla Katiraee, Dr. Andrew Carson, Anath Lionel and Matt Gazzellone, I like to thank you for sharing your experience, and for listening to my frustrations. I like to thank Dr. Christian Marshall, Dr. Richard Wintle, Bhooma Thiruvahindrapuram, and Dr. Sergio Pereira for their help and for sitting through many of my practice presentations. I like to thank my collaborators Dr. Samuel Levy, Dr. Ewen Kirkness, Dr. Pauline Ng, Dr. Vanessa Hayes, Dr. Matt Hurles, Dr. Donald Conrad, Dr. Charles Lee, Dr. Hansoo Park, Dr. Timothy Harkins, Dr. Clarence Lee and Dr. Brian O'Connor who made the findings that make this thesis possible. I greatly appreciate the support given to me by the various funding agencies: National Sciences and Engineering Research Council of Canada, University of Toronto and The Hospital for Sick Children. I'd like to thank all past and present members of TCAG and the Scherer lab for their kindness and help, Sanjeev Pullenayegum, Kozue Otaka, Karen Ho, Jessy Lyons, Jenny Kaderali, Thomas Nalpathamkalam, just to name a few. Last but not least, I like to thank my parents for their love and continual support during this long, rewarding and sometimes challenging journey.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER I: INTRODUCTION TO STRUCTURAL VARIATION

I.A Variation in the human genome

The human genome is comprised of some six billion nucleotides of information packaged in 23 sets of inherited chromosomes. A striking observation from studying the human genome is the extent of similarity among individuals across populations. Therefore, we can gain insights of evolution, human diversity, and disease susceptibility by studying a small fraction of the genome that is variable between people.

There are many forms of genome variation. Single nucleotide variants are substitution changes in DNA sequence, and are the most common form of variation. Those with allele frequency of over 1 % are called single nucleotide polymorphisms or SNPs. Structural variation refers to cytogenetically visible rearrangements, and more common submicroscopic variants, including deletions, insertions, and duplications – collectively termed copy number variation (CNV) – and inversions and translocations. CNVs are defined as gain or loss of DNA fragments whose length is 1 kilobases (kb) or above, and those below that size are usually classified as small insertions/deletions (indels) (Feuk, et al., 2006a). In general, structural variation is less numerous in number than SNPs, but has a wider size distribution.

Earlier discoveries of human genetic variation are made from clinical studies. In 1949, Pauling and colleagues found a difference in electrophoretic mobility of hemoglobin derived from erythrocytes of normal individuals and from sickle cell anemic samples. They termed sickle cell disease as a molecular disease (Pauling, et al., 1949). The first chromosomal disease was reported in 1959, when trisomy of chromosome 21 was found in individuals with Down Syndrome (Lejeune, et al., 1959). In the same year, sex chromosome anomalies were discovered in people with Klinefelter Syndrome and Turner Syndrome (Ford, et al., 1959). Two other cases of aneuploidies, trisomies 13 (Patau, et al., 1960) and trisomies 18 (Edwards, et al., 1960) were described in 1961. Another form of cytogenetic change is heteromorphisms, which are large, microscopic visible chromosomal rearrangements. In 1961, Novell and Hungerford reported the Philadelphia chromosome abnormality associated with chronic myelogenous leukemia (Nowell and Hungerford, 1961). This is the translocation between chromosome 9 and 22. Finally, microscopically visible fragile site at

Xq27.3 (Lubs, 1969), associated with X-linked mental retardation, is caused by massive expansion of triplet repeat in the *FMR1* gene (fragile X mental retardation 1).

On the other end of the size spectrum at the submicroscopic level, single nucleotide variants and small structural variants can disrupt genic or regulatory sequences and cause single-gene disorders. The discovery of these mutations was made possible by the inventions of cloning (Cohen, et al., 1973) and DNA sequencing (Maxam and Gilbert, 1977; Sanger, et al., 1977). In 1977, Frederick Sanger developed DNA sequencing based on chain termination method, and this technique is now commonly known as Sanger sequencing or the "first generation sequencing".

Duchenne Muscular Dystrophy is a severe X-linked disorder. The *DMD* gene is one of the largest protein-coding genes in the human genome, and it was cloned in 1987 (Koenig, et al., 1987). Mutation in this gene include large deletions (60-65%), large duplications (5-10%), small indels or substitutions (5-10%), and in rare cases, X-autosome translocations in females (Nussbaum, et al., 2007; Ray, et al., 1985). In 1989, the cystic fibrosis gene was mapped (Kerem, et al., 1989; Riordan, et al., 1989; Rommens, et al., 1989). We now know that there are over 1,930 mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*), ranging from missense mutations to large indels (http://www.genet.sickkids.on.ca/app). The Huntington Disease gene was subsequently cloned in 1993. The unstable expansion of trinucleotide repeats is associated with this autosomal dominant disorder (MacDonald, et al., 1993).

The availability of the human reference genome assembly since 2001 is another major resource for variation studies. Not only is it the map for genes and other functional elements, it is also the reference upon which genomic variants are defined: Its sequence is used to design microarray probes, and to which sequenced DNA is mapped. During the course of determining the human genome sequence (Lander, et al., 2001; Venter, et al., 2001), and subsequently through the International HapMap Project (The International HapMap Consortium, 2005), millions of single nucleotide variants were identified, and were determined to be the most plentiful form of variation in the genome. Overall, single nucleotide variation accounts for the polymorphisms that mark the widespread genetic

diversity within our species, and also account for the mutations that contribute to disease. It was estimated that 0.1 % of the genome is different between any two human beings due to SNPs (Lander, et al., 2001; Venter, et al., 2001).

I.B Prevalence of structural variation

While cytogenetic rearrangements are rare and tend to be associated with disease phenotypes, SNPs are common and are observed in all individuals. Yet, information about variation between chromosomal rearrangements and nucleotide substitutions has been less extensive. The developments of advance genome-wide scanning technologies (Iafrate, et al., 2004; Redon, et al., 2006; Sebat, et al., 2004) and DNA sequence comparative analyses (Feuk, et al., 2005; Khaja, et al., 2006; Tuzun, et al., 2005) enable investigation of the extent of structural variation in the general population.

A common theme in structural variation studies is that structural variation is prevalent in the genome, and is present in all individuals. Particularly, in an early study, Redon and colleagues discovered 1,447 CNVs, covering a remarkable 12 % of the human genome (Redon, et al., 2006). Subsequently, with higher resolution and greater precision microarrays, that number has been refined to about 3.7 %, which is about 0.7 % when comparing two genomes (Conrad, et al., 2010b). Evidently, the greatest source of genetic diversity in humans lies not in SNPs (contributing to 0.1 % of diversity) but rather in larger structural variants (Conrad, et al., 2010b; McCarroll, et al., 2008; Redon, et al., 2006).

I.B.1 Methods of discovery

In recent years, there had been development of many genome-wide experimental and computational strategies to detect structural variation, and here I discuss the ones that are of the greatest impact to variation discovery. The exploration of submicroscopic genomic unbalanced structural variants was made possible with the development of microarray technology. High resolution array comparative genome hybridization (CGH) and SNP-genotyping microarrays are the most widely used approaches to detect CNVs (Conrad, et al., 2010b; McCarroll, et al., 2008; Redon, et al., 2006). In the case of array CGH, the probe fluorescence signal ratio between a test and a reference sample acts as proxy for copy number. A rise in ratio represents a gain in copy number in the test sample with respect to the

reference, and conversely a drop represents a loss in the test sample. Consecutive probes with aberrant signal ratio are required to call a CNV. Thus the resolution of microarrays is determined by both the number of probes used to make a call and the spacing between adjacent probes.

The principle of genotyping arrays is similar except the probe signal intensities of each sample are compared against a reference consisting of a collection of sample hybridizations (Alkan, et al., 2011). So, hybridization intensities are compared with average values derived from the reference, and any deviation from the averages indicates copy number change. SNP arrays provide both copy number information and genotype information. For instance, they can identify loss of heterozygosity, which can support the presence of a deletion or segmental uniparental disomy.

Besides microarrays, sequence-based approaches can also detect genomic variation. Typically, DNA sequence reads of a test individual generated by whole-genome shotgun sequencing is aligned to the National Center for Biotechnology Information (NCBI) reference assembly, and discordant signatures of the alignments would indicate variants. Four major computational approaches have been applied to sequence data to detect genomic structural variation, and they are the mate-pair, split-read, read depth and assembly comparison approaches. These approaches are initially developed based on Sanger sequences (Sanger, et al., 1977), but they can also be adopted to work on Next Generation Sequencing (NGS) data.

Paired sequences (a mate-pair) from an insert library of defined sizes created from genomic DNA from a test individual are sequenced, and then aligned to the reference assembly. The distance between the pair is then compared with the expected size of the insert. Any discrepancy in distance between the observed and expected size indicates putative insertion or deletion, while any incorrect orientation with respect to the reference assembly highlights potential inversion (Hormozdiari, et al., 2009; Kidd, et al., 2008; Korbel, et al., 2007; Tuzun, et al., 2005). Secondly, small indels can be captured by identifying intra-alignment gaps or “split-reads” within the alignment of sequence traces (Mills, et al., 2006; Ye, et al., 2009). A third type of variant-calling approach is the read depth approach, whose idea is that

duplicated regions will show significantly elevated number of alignments compared to diploid regions, whereas deletions will show reduced levels of alignments (Abyzov, et al., 2011; Chiang, et al., 2009). Finally, structural variants can also be identified by aligning two assemblies of DNA. Here, the mate-pair reads from the test sample can be assembled *de novo* into long range scaffolds, which are then subsequently mapped to the public reference assembly. In principle, this assembly comparison method can identify all types of variants: SNPs, insertions, deletions, duplications, inversions and translocations (Feuk, et al., 2005; Khaja, et al., 2006; Levy, et al., 2007; Li, et al., 2010b; Zerbino and Birney, 2008). Furthermore, it can resolve the structures of regions where multiple events having occurred in close proximity, and in addition to identify a duplication event, it can pinpoint the location of the duplicated copy.

I.B.2 Structural variation: type, number, size and detection platform

Table I. 1 shows substantial differences in type, number and size of structural variants among studies. Difference in platform, probe density, and computational algorithm can yield different number of CNVs that are of different size (Alkan, et al., 2011; Scherer, et al., 2007). There are a few general characteristics common in array-based studies. More deletions are detected than duplications. Copy number changes of repetitive elements are usually not reported because either these loci cannot be uniquely targeted by short oligonucleotide probes, or over-saturation of probe fluorescence would prevent an accurate high copy number count. Also, arrays cannot detect dosage-invariant changes such as inversions and translocations.

Table I. 1. Structural variation and detection methods.

Method	Samples	Deletions		Insertions		Inversions		Duplications		Reference
		#	Median (bp)	#	Median (bp)	#	Median (bp)	#	Median (bp)	
SNP array[1]	270	1,122	6,216	-	-	-	-	442	14,122	(McCarroll, et al., 2008)
Array CGH[2]	40	7,909	2,284	-	-	-	-	4,740	5,265	(Conrad, et al., 2010b)
Array CGH[3]	30	14,597	2,439	-	-	-	-	5,502	3,835	(Park, et al., 2010)
Mate-pair	8	1,843	8,657	560	7,594	1,146	77,119	1,768	8,429	(Kidd, et al., 2008)
Split-read[4]	36	216,212	2	199,222	2	-	-	-	-	(Mills, et al., 2006)

[1] Affymetrix 6.0 SNP. [2] NimbleGen 42M oligonucleotide array set. [3] Agilent 24M oligonucleotide array set. [4] Unique indels only.

Alignment of mate-pair sequence can detect more deletions than insertions (Kidd, et al., 2008). Whereas there is no theoretical upper size limitation in detecting deletions, insertion identification is restricted by the length of insert fragment, thus limiting the number of insertions detectible. In addition, the DNA composition of the insertion is typically unknown as only the ends of insert fragments are sequenced. An advantage of mate-pair mapping approach over microarrays is that it can detect inversions (Table I. 1).

Mills and colleagues used the split-read approach to call structural variation (Mills, et al., 2006). While they called the most number of variants, but their size is much smaller than the other studies (Table I. 1). Similar to mate-pair mapping, the split-read approach calls fewer insertions than deletions, because of insertions are bounded by the size of sequenced reads. This method also cannot detect any copy balanced variation.

In summary, different platforms can identify different type, number and size of structural variants. This is a challenge for structural variation. If a study uses only one detection approach, it will miss certain types of variation. As describe in Chapter II, in order to compensate for the shortcomings of each method, I used multiple detection methodologies to identify structural variation in a human individual. By examining the same sample using multiple methodologies, I quantified the type and size range of variants that can be detected by each methodology, and most importantly, showed that presently there is no single method that can readily capture all variation (Pang, et al., 2010)

I.B.3 Genomic impact

Structural variation is ubiquitous in the genome. The Database of Genomic Variants (DGV) is a repository for structural variation found in general population surveys (Iafrate, et al., 2004; Zhang, et al., 2006). As of September 2012, DGV has 833,981 gains and losses and 906 inversion entries, covering over 30 % of the euchromatic region of the genome. Figure I. 1 shows the size distribution of the gains, losses and inversions contained in DGV.

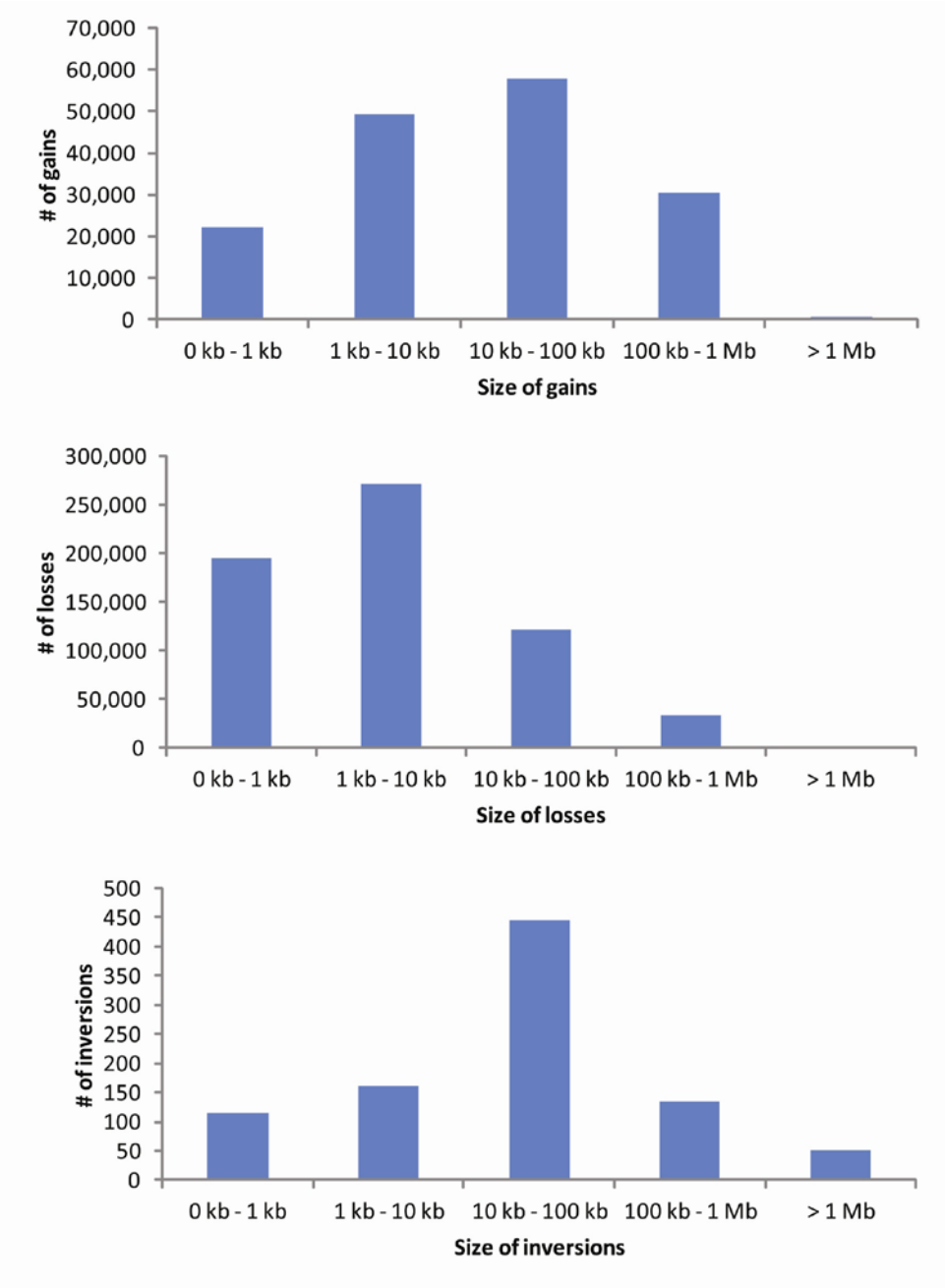


Figure I. 1. Size distribution of gains and losses, and inversions in DGV.
(Top) The size distribution of insertions and duplications and (Middle) deletions reported in DGV. The 0 kb - 1 kb size range is likely an underrepresentation. That size range is below the detection limit of most microarrays, which are currently the most used CNV-detection method. There are more deletion records than insertions and duplications. Although small variants are under-represented in both (A) and (B), the trend of deletion is probably more representative of variability in the genome than insertion or duplication. (Bottom) The size distribution of inversions, and it is discussed in more details in text.

When deletions or duplications encompass genic or regulatory regions, they may create an imbalance in the appropriate level of RNA or protein produced. Furthermore, when CNVs overlap genes encoding transcription factors, they can affect the expression of both the transcription factors and their targeted genes. For genes or pathways that are dosage sensitive, CNVs may lead to change in gene-dosage, thus susceptibility to disease. Duplication events can mediate additional structural variants by catalyzing non-allelic homologous recombination (NAHR) between the duplicated copies. In addition, gene duplications can create redundancy such that the new copies can accumulate mutations, thus allowing them to gain new molecular functions. Alternatively, if the breakpoints of duplications, deletions, inversions, and translocations partially overlaps genes, then the variants can alter the genes' structure by disruption or by creation of novel transcripts (Feuk, et al., 2006a).

Hundreds of structural variants, mostly CNVs, can be detected in any individual using microarrays or sequencing. A study by our group reported 3,340 CNVs overlap 2,698 RefSeq genes, altering the structure of 3,863 transcripts and 1,519 coding sequence. In general, though, there is a paucity of CNVs overlapping genes, and the impoverishment of deletion is stronger than duplication, perhaps due to more severe consequence associated with loss-of-function (Conrad, et al., 2010b). Copy number variable regions show increased SNP variation and a higher density of short genes, while regions that are stable against CNVs are enhanced for longer genes and ultra-conserved elements (Johansson and Feuk, 2011). Of the genes impacted by CNVs, those that are involved in interactions with the environment such as immunity or sensory perception are enriched. On the other hand, genes that are involved in basic development such as the development of nervous system are typically dosage sensitive, and are enriched in "CNV deserts", thus suggesting that variation may lead to reduced organismal fitness.

To estimate the relative genomic impact of substitution variation and structural variation, one can examine data from disease and known mutations. Table I.2 shows the relative impact of different types of mutations underlying characterized disease phenotypes as reported in the Human Gene Mutation Database (Table I. 2). According to this, missense and nonsense

substitutions account for the majority of reported mutations. However, there are a few caveats. First, reported mutations involved in genetic disease tend to reside in the coding regions. Although well characterized, coding regions constitute only a small portion of the genome. Hence, the proportions listed here may not be representative of the whole genome. Furthermore, it should be noted that substitutions are easier to detect than structural variants, so the effects of the latter would also be less well annotated. Finally, in terms of number, there are more substitutions reported to be associated with disease phenotypes. But in terms of size, one large rearrangement can in principle overlap multiple neighbouring genes, affect multiple physiological pathways, and potentially have more serious effects than single base change. Therefore due to these confounding factors, it is perhaps premature to draw conclusions on the relative genomic impact of substitutions and structural variants.

Table I. 2. Relative frequency of different types of mutations underlying disease phenotypes.
(data from the Human Gene Mutation Database, March 2013)

Mutation type	Number	% of Total
Missense/nonsense substitutions	74,328	55.5
Splicing substitutions	12,414	9.3
Regulatory substitutions	2,628	2.0
Small deletions	20,705	15.4
Small insertions	8,558	6.4
Small indels*	1,987	1.5
Gross deletions	9,479	7.1
Gross insertions	2,172	1.6
Complex rearrangements	1,341	1.0
Repeat variations	407	0.3
Total (in 5,283 genes)	134,019	100

* Co-localised insertions and deletions

I.B.4 Structural variation and selection

From population studies, CNVs tend to bias away from functional sequences (Conrad, et al., 2010b; Redon, et al., 2006). As evident by the depletion of structural variation, purifying selection acts most strongly on exonic, then intronic, and then intergenic sequences (Conrad, et al., 2010b; Mills, et al., 2011a; Redon, et al., 2006). Also, the genome is more tolerable of submicroscopic duplications than deletions; there is a greater proportion of full-gene or exonic duplication than deletion (Conrad, et al., 2010b). Furthermore, ultra-conserved elements, which are perfectly conserved between orthologous regions of human, mouse and rat (Bejerano, et al., 2004), are known to be under intense purifying selection. These sequences are significantly depleted within CNVs. Amplification or deletion of these critical regions may result in deleterious consequences. All these observations substantiate the presence of negative selection can act on structural variation.

There is also evidence of positive selection on structural variants. Extended long range haplotype around a variant is indicative of positive selection (Sabeti, et al., 2002). A deletion between pathogen immunity genes *APOL2* and *APOL4* reside in a long extended haplotype (Genovese, et al., 2010). These genes have been shown to be under positive selection in primates. Population differentiation can also identify variants which are differentially selected for in different environments, so it is also indicative of positive selection acting in one or more populations. The copy number of the amylase *AMY1* gene exhibits high population differentiation. The gene is responsible for the hydrolysis of starch. Populations, such as Japanese, that have high consumption of starch tend to have higher copy number of AMY1 than populations that consume starch in low quantity, such as Biaka pygmy. This situation is believed to be positive selection in response to diet (Perry, et al., 2007).

I.B.5 Mechanisms of structural variation formation

The formation mechanisms of structural variation can be inferred by examining the underlying DNA sequence. Highly homologous duplicated sequences, such as segmental duplications (operationally defined as duplicated sequences > 1 kb sharing over 90 %

sequence identity) or retrotransposable elements, can mediate NAHR. Tandem homologous sequences can lead to the formation of gains and losses (Conrad, et al., 2010b; Kidd, et al., 2008), oppositely oriented ones can mediate inversions (Feuk, et al., 2005), and homologs situated on different chromosomes can cause translocation (Ou, et al., 2011). On the other hand, rearrangements that lack significant flanking sequence homology or that show short stretches of flanking microhomology are believed to be formed by ligation processes such as nonhomologous end-joining (NHEJ) or microhomology-mediated end-joining (MMEJ). Complex genomic rearrangements consist of more than one simple rearrangement, and have two or more breakpoint junctions. These variants are then proposed to have been formed by either strand slippage or template switching at a replication fork that is stalled (fork stalling and template switching, FoSTeS) (Lee, et al., 2007) or broken (microhomology-mediated break-induced repair, MMBIR) (Hastings, et al., 2009).

Small indels can be simple additions or deletions of bases due to errors in replication or repair. To the contrary, indels associated with tandem repeat are believed to be due to errors in replication and recombination. While small mutations involving gain or loss of < 10 repeats are believed to be caused by replication slippage, large mutations are compatible with recombination (Richard and Paques, 2000). Finally, three classes of retrotransposable elements – L1 (long interspersed element 1), SVA (short interspersed element (SINE-R), variable number of tandem repeats (VNTR), and Alu), and Alu (a SINE) elements – are still active in the human genome. These DNA retrotransposons are transcribed into RNA intermediates, reverse transcribed, and finally randomly re-inserted themselves in distal locations (Konkel and Batzer, 2010).

One can infer the origin of structural variation by investigating the DNA sequence of the variant or its surrounding region. For example, the presence of long homologous segmental duplication flanking a deletion would suggest the deletion is formed by NAHR, whereas an insertion of DNA resembling an Alu would indicate a retrotransposition event. A prerequisite to this sequence-based inference of mutational mechanism is to have precise breakpoint; however, this information is not easy to obtain. Spacing between microarray probes prevents accurate delineation of precise variant breakpoint; the true breakpoint can reside in DNA sequence between neighbouring probes. Short-read sequencing platforms similarly generate

imprecise variant boundaries, as evident in the recent 1000 Genomes Project structural variation study, where only half (53%) of the calls have been mapped at nucleotide level (Mills, et al., 2011b). Furthermore, existing mechanism studies show various results in the proportion of mechanistic process (Table III. 1), and this discrepancy is due to the bias in examining only subsets of variations in the human genome (for example, only the indels or only variants of specific sizes). In Chapter III of this thesis, I provide a more accurate estimate of the proportion of various mutational processes by annotating the sequence content of a near-complete set of breakpoint-refined structural variation discovered in a fully-sequenced human genome.

I.B.6 Inversions

Because inversions are generally copy number invariant, they often have no functional significance unless their breakpoints disrupt genes or fall between genes and their transcription regulatory elements. Examples where recurrent inversions have been shown to lead to disruption of genes thus leading to clinical phenotypes are the disruption of factor VIII gene in hemophilia A (Green, et al., 2008), iduronate 2-sulphatase gene in Hunter syndrome (Bondeson, et al., 1995), emerin gene in Emery-Dreifuss muscular dystrophy (Small, et al., 1997). Other inversions associated with phenotype but not directly causative are the ones that increase the risk of further rearrangements. In some microdeletion syndromes such as the Williams-Beuren Syndrome (Osborne, et al., 2001), phenotypically normal parents of patients have been shown to carry an inversion in the deletion interval, and that the inversion is observed in the parent who transmits the disease-related chromosome. Generally, the mechanism of how an inversion mediates subsequent rearrangements is still not well understood.

Contrasting with gains or losses of DNA, which can be detected by commonly-used microarrays, whole-genome sequencing is required to directly detect novel genomic inversions. Mate-pair mapping is the only practical approach currently applied to detect submicroscopic inversions genome-wide. Therefore, due to the difficulty in their detection, the map of inversion in the human genome is far lagging behind gains and losses.

There are 833,981 CNV, but only 906 inversion entries in the DGV as of September 2012 (Iafrate, et al., 2004; Zhang, et al., 2006). Nonetheless, by examining the size distribution of variants reported in DGV, one would notice that most inversions are generally in the 10 to 100 kb size range, contrasting to deletions in the 1 to 10 kb size range (Figure I. 1). There are different possible explanations to the difference in size distribution. First, large inversions may be less detrimental than large CNVs. Functional sequences within the inverted region are essentially unchanged, but those inside CNVs are always changed in copy number. Hence, large inversions are less likely to cause phenotypic change than large CNVs. For instance, even cytogenetically visible inversions can have no visible phenotypical effect (Nussbaum, et al., 2007). On the other hand, the difference in size distribution may also be due to methodological limitations. Since the majority of the inversion records in DGV are discovered by mate-pair mapping, the size distribution may simply correspond to the resolution of the approach. There may be additional inversions, particularly at the small size range, remaining to be detected.

I.C Personal genome sequencing

New sequencing technologies significantly improve the capability to perform whole genome sequencing. There are three main differences between NGS and Sanger sequencing: parallelization, high throughput and reduced cost. Consequently, whole genome sequencing is now more affordable and can be completed in a shorter time span. For example, the first personal genome sequenced (the genome of J. Craig Venter, also called the HuRef genome) (Levy, et al., 2007), which used Sanger-based capillary sequencing technology, cost approximately 2 million US dollars (Stepanov, 2010). Subsequently, the genome sequences of Yoruba and Chinese individuals cost about $250,000 and $500,000, respectively (Bentley, et al., 2008; Wang, et al., 2008). The trend is dramatically decreasing with cost drops to about $4,400 (Drmanac, et al., 2010). In return for efficiency, the data generated by NGS tend to have higher error rates than Sanger approach (Liu, et al., 2012), although that is rapidly improving. The fragments or inserts used for end-sequencing (~150 bp to ~2 kb) are usually much shorter than the recombinant DNA clones used in sequencing the HuRef genome (2 kb to 37 kb). Finally, NGS sequenced DNA traces (25 bp to ~ 200 bp) are typically much shorter than Sanger reads (average of ~700 bp). There are several NGS

platforms that have been used for human genome sequencing, and they are Roche 454, Illumina, Life Technologies SOLiD, Helicos, and Complete Genomics and Life Technologies Ion Torrent.

Hundreds of human genomes – controls or disease cohorts – have been sequenced to date. The ultimate goal of these sequencing projects is to capture the full spectrum of genetic variation that will facilitate medical interpretation. Different sequencing platforms with vastly different chemistries have been used, and all except the HuRef genome, have been generated by NGS (Table I. 2). It is important to note here that the NCBI and Celera assemblies generated since 2001 consist of a mosaic of haploid DNA derived from multiple individuals, and hence are not considered as personal genomes.

Table I. 3. Summary of personal sequencing studies.

Date	Sample	Pop.	Platform	Cov	SNP	Gain/loss*			Inversion*			Reference
						#	Min size (bp)	Max size (kb)	#	Min size (bp)	Max size (kb)	
**2007, Oct.	Venter (HuRef)	Caucasian	ABI3730xl	7.5	3,213,401	796,079	1	82.7	90	120	686.3	(Levy, et al., 2007)
2008, Apr.	Watson	Caucasian	454	7.4	3,322,093	222,718	2	38.9	0	0	0	(Wheeler, et al., 2008)
2008, Nov.	NA18507	Yoruba	Illumina	41	4,139,196	410,120	1	50.0	0	0	0	(Bentley, et al., 2008)
2008, Nov.	Yanhuang (YH)	Chinese	Illumina	36	3,074,097	137,927	1	180.0	17	282	158.3	(Wang, et al., 2008)
2009, May	Seong-Jin Kim (SJK)	Korean	Illumina	29	3,439,107	345,885	1	99.5	415	100	98.0	(Ahn, et al., 2009)
2009, Jun.	NA18507	Yoruba	SOLiD	17.9	3,866,085	232,124	1	97.0	91	112	90.5	(McKernan, et al., 2009)
2009, Aug.	AK1	Korean	Illumina	27.8	3,453,653	171,439	1	3,675.8	0	0	0	(Kim, et al., 2009)
2009, Aug.	Quake (P0)	Caucasian	Helicos	28	2,805,471	752	N/A	N/A	0	0	0	(Pushkarev, et al., 2009)
2010, Jan.	NA07022	Caucasian	Complete Genomics	87	3,076,869	337,635	1	50bp	0	0	0	(Drmanac, et al., 2010)
2010, Jan.	NA19240	Yoruba	Complete Genomics	63	4,042,801	496,194	1	50bp	0	0	0	(Drmanac, et al., 2010)
2010, Jan.	NA20431	Caucasian	Complete Genomics	45	2,905,517	269,794	1	50bp	0	0	0	(Drmanac, et al., 2010)
2010, Feb.	Saqqaq	Paleo-Eskimo	Illumina	20	2,193,396	0	0	0	0	0	0	(Rasmussen, et al., 2010)

2010, Feb.	KB1	Khoisan	454/Illumina	33.4	4,053,781	463,788	1	93.3	0	0	0	(Schuster, et al., 2010)
2010, Feb.	Tutu (ABT)	Bantu	SOLiD/Illumina	37.2	3,624,334	3,395	1	11bp	0	0	0	(Schuster, et al., 2010)
2010, Mar.	Pedigree #1 mother	Caucasian	Complete Genomics	51	~2,900,000	N/A	N/A	N/A	0	0	0	(Roach, et al., 2010)
2010, Mar.	Pedigree #1 father	Caucasian	Complete Genomics	88	~3,200,000	N/A	N/A	N/A	0	0	0	(Roach, et al., 2010)
2010, Apr.	Lupski	Caucasian	SOLiD	29.9	3,420,306	234	1,690	1,627.8	0	0	0	(Lupski, et al., 2010)
2010, Sep.	Irish	Irish	Illumina	11	3,125,825	195,798	1	29	0	0	0	(Tong, et al., 2010)
2010, Nov.	NA18943	Japanese	Illumina	40	3,132,608	5,319	1	221.8	57	N/A	N/A	(Fujimoto, et al., 2010)
2011, Jul.	Moore	Caucasian	Ion Torrent	10.6	2,598,983	3,391	50	982.8	22	250	1,941.9	(Rothberg, et al., 2011)
2011, Oct.	Aboriginal Australian	Aboriginal Australian	Illumina	6.4	449,115	22,576	N/A	N/A	0	0	0	(Rasmussen, et al., 2011)
2012, Feb.	Tyrolean Iceman	Southern European	SOLiD	7.6	2,218,163	N/A	N/A	N/A	0	0	0	(Keller, et al., 2012)
2012, Mar.	Michael Snyder	Caucasian	Complete Genomics/ Illumina	150/ 120	3,301,521	219,342	1	>50bp	N/A	N/A	N/A	(Chen, et al., 2012)
2012, Jul.	IGIB1	Indian	Illumina	28	3,409,125	491,119	N/A	N/A	49	N/A	N/A	(Patowary, et al., 2012)
2012, Aug.	SAIF	Indian	Illumina	34.9	3,459,784	384,926	1	335bp	0	0	0	(Gupta, et al., 2012)
2012, Oct	NA18507	Yoruba	Illumina	41	0	785,077	1	9,214.9	172	235	8,749.4	(Jiang, et al., 2012)

* N/A denotes that data has been detected and published, but the details on the number or size are not available in the study. 0 means no data of this type was detected in the study.

** From the HuRef study, I include all homozygous indels, heterozygous indels, indels embedded within simple, bi-allelic, and non-ambiguously mapped heterozygous mixed sequence variants, and only those inversions whose size is at most 3Mb.

J. Craig Venter is a pioneer in genome research, and is best known for his work in using whole-genome shotgun sequencing approach to generate a draft of the human genome in 2001(Venter, et al., 2001). Besides the sequencing of his own personal genome in 2007 (Levy, et al., 2007), he led a team to construct the first synthetic bacterial cell in 2010 (Gibson, et al., 2010). He is the founder of Celera Genomics, The Institute for Genomic Research, the J. Craig Venter Institute and Synthetic Genomics (Venter, 2007). In the beginning of my PhD study in 2007, I participated in the sequencing and characterization of the Venter genome, herein called the HuRef genome. The HuRef assembly was produced from ~ 32 million random paired clone-end DNA fragments. Long and high quality reads (average of ~ 700 bp) sequenced from the ends of long clone fragments of multiple insert lengths (2 kb, 10 kb and 40 kb) enabled the generation of *de novo* long-range assembly. This HuRef diploid assembly was then compared to the NCBI public reference assembly and revealed 3,213,401 SNPs, 292,102 heterozygous indels, 559,473 homozygous indels, and 90 inversions. Structural variation accounts for 74 % of variable bases. The study reveals that there are ~ 12,500 non-silent coding variants in the HuRef genome. Particularly, he is heterozygous for alleles that are linked to susceptibility to coronary artery disease, hypertension and myocardial infarction. However, the majority of coding variants in his genome are neutral or nearly neutral, and that there is no evidence of severe disease (Ng, et al., 2008).

Besides the HuRef individual, other genomes of controls from different populations have been sequenced (Table I. 2). These studies use different NGS technologies, achieve different levels of coverage with different amounts of variation in different parts of the genome. The consensus of the results is that a genome has about 3.2 million SNPs; however, there is no agreement on the number of structural variants. The number of gains and losses reported is variable, and inversions are rarely reported at all. This again exemplifies the difficulty in detecting structural variation, which unlike SNPs, many of them cannot be captured within short sequence reads.

Overall, these personal genome studies show that there is a tremendous amount of variation in the human genome, but also a limited understanding of the effect of most of these calls.

Although numerous variants are discovered in each genome, only a minority of which can correlate with the physical trait of the individual. Additional studies, such as the 1000 Genomes Project, are needed to sequence a large number of individuals to catalog genomic variation in humans (Durbin, et al., 2010).

I.D Application of whole genome sequencing

Some benefits of genomic sequencing are already apparent, particularly in diagnosis in identifying variants associated with monogenic disease. For example, Choi and colleagues reported sequencing of the exome – the coding regions – of a patient to diagnose for possible Bartter syndrome, a disease of problem in salt re-absorption (Choi, et al., 2009). Since the healthy parents were first cousins, the authors searched for loss of heterozygosity, and identified a novel homozygous missense mutation in the SLC26A3 gene in 7q31.1. This gene is known to cause congenital chloride-losing diarrhea (OMIM 214700). Then clinical follow-up found that the patient indeed had chloride-losing diarrhea, which was not initially considered, and not Bartter syndrome. Also, Lupski and colleagues sequenced the genome of James Lupski, who has Charcot-Marie-Tooth neuropathy (Lupski, et al., 2010), using NGS achieving an average depth of ~ 30 X. After focusing on 40 candidate neuropathy genes, the authors identified two mutations that are compound heterozygous at the SH2 domain and tetratricopeptide repeats 2 gene (*SH3TC2*). Interestingly, subsequent examinations of other family members – all being compound heterozygotes for these variants – showed that each heterozygous variant co-segregated with an electrophysiological phenotype such as axonal neuropathy or carpal tunnel syndrome. Finally, a recent study demonstrates that diagnosis in neonatal intensive care units by whole genome sequencing can be achieved in 50 hours. Because of the rapid course of monogenic diseases, genetic heterogeneity, and that current tests only identify a few disorders at a time, existing newborn screens of acutely ill neonates may not be made in time. Short turn-around time of sequencing result, coupling with symptom-assisted analysis, can potentially shorten diagnosis and quicken clinical decision making (Saunders, et al., 2012).

Another application of genome sequencing is on the studying of population genetics. In principle, whole-genome sequencing offers unprecedented resolution to detect of all genomic variants along the chromosomes, so sequencing of parents and children can directly

determine the rate of mutation. Conrad and colleagues have used sequencing to examine the *de novo* rate of single nucleotide variants, and show a mutation rate about $1.1x10^{-8}$ per nucleotide per generation, equivalent to ~ 70 new mutations in the genome per generation (Conrad, et al., 2011). Moreover, studying recombination rate in people once seemed impossible, unless one can find individuals with hundreds of children and then sequence their genomes. Advancement in single-cell sequencing enables examination of recombination rate in spermatogenesis. Wang and colleagues have isolated, and sequenced 100 single sperm genomes (Wang, et al., 2012). They observed recombination rate at an average of 22.8 events per cell, identified recombination hotspots, and estimated gene conversion rate at 5 – 15 per cell. The location of recombination hotspots can potentially help future studies to identify hotspots of structural variants.

Unlike SNP-microarrays that are designed to target common variants, sequencing can capture both common and rare variants. Recent studies have examined the frequencies of SNPs detected, and identified an excessive number of low frequency rare variants in cohort samples. This is due to the explosive accelerated growth of population size in recent 100 generations (Coventry, et al., 2010; Gravel, et al., 2011; Keinan and Clark, 2012). Rapid growth increases the load of rare variants, as there is little time for natural selection to operate and remove them, unless they are severely deleterious. The studies suggest that these variants may play a role in the genetic burden of complex disease risk, and that future disease studies will need a large sample size to capture these individually rare events.

Finally, one can also make use of the data generated from sequencing to design additional assays to genotype structural variants, which are more difficult to detect than SNPs. Based on indels discovered in sequencing studies, Mills and colleagues designed a custom microarray to genotype over 10,000 of these variants in a panel of individuals (Mills, et al., 2011a). They characterized their allele frequencies and inheritance patterns. They also found high linkage disequilibrium (LD) of indels with SNPs in the HapMap project, thus enabling potential integration of the indels to existing haplotype map of the human genome. Similarly, I leveraged the availability of the breakpoints of inversions in HuRef, and developed assays to genotype a subset of these variants in multiple subjects. I found that submicroscopic

inversions may not truly be copy-balanced, may have net change in DNA content, and can have complex structures that may lead to reference genome misassembly.

I.E Annotation of structural variation in a personal genome sequence

As mentioned above, there are still many existing challenges to detect and characterize structural variation. The rationale of my project is that better analysis tools and a deeper understanding of genomic variation are essential to decipher and interpret the human genome. I have three main objectives: 1) develop methods to analyze genomic data and detect structural variation; 2) annotate the formation mechanisms of structural variation; and 3) genotype submicroscopic inversions in human populations. The primary DNA sample used for my experiments is the HuRef sample, because of availability of DNA, full-access to sequence data, and known information on the identity and phenotype of the donor.

In the original Levy et al. study, the HuRef variation set was generated by comparison of the HuRef assembly with the NCBI public reference assembly Build 36 (Levy, et al., 2007). In principle, the assembly comparison method can detect all types of variation: substitutions, insertions, deletions, duplications and inversions. This approach depends on the length and quality of the assembled sequence scaffolds (Levy, et al., 2007). Of the deletions called by CGH and SNP microarray experiments run with the HuRef DNA, interestingly, none of them had been detected by the sequence-based assembly comparison approach. In addition, there is an under-representation of heterozygous indels due to the relatively low coverage achieved; about 44 to 52 % of the heterozygotes are estimated to have been labeled as homozygotes. Hence, the variation map of the HuRef genome is not complete.

This thesis describes my effort in building what is currently the most detailed variation map of a human genome. It has three chapters, and each addresses an outstanding problem in variation studies.

Chapter II: *Towards a comprehensive structural variation map of an individual human genome.* I used a combination of computational and experimental approaches to identify additional structural variation missing in the initial Levy et al. study. First, I implemented mate-pair and split-read algorithms to detect insertions, deletions and inversions by sequence

read alignments. Next, I found novel CNVs generated from high-density CGH and SNP microarrays. The large structural variation detected by my multi-platform approach complements with the smaller size variants found by assembly comparison. My work demonstrated that variant discovery is largely dependent on the strategy used, and presently there is no single method that can readily capture all types of variation and that a combination of strategies is required. The results described therein provide a foundation to the analysis in subsequent chapters.

Chapter III: *Mechanisms of formation of structural variation in a human genome.* There are a few genome-wide studies examining the mutational mechanism underlying the formation of structural variations, and their data show varying results in terms of the relative proportion of contributing mechanisms. In this chapter, I provide a thorough annotation of formation mechanism of structural variation in the HuRef genome. Leveraging the availability of precise junction information in the long-read Sanger sequences, I inferred the formation mechanism for the entire size spectrum of structural variation. With this unique data, I discovered that different mechanisms are more prominent within different size classes of variants. Comparing my data to other published results, I showed that a large number of variants had previously been overlooked, with noticeable gaps in annotation at specific variant size.

Chapter IV: *Complex breakpoint structures associated with microscopic inversions.* Inversion discovery is rather modest compared to copy number changes, mostly due to the limited number of high-throughput genomic tools. Accurate breakpoint information from HuRef variants offers an opportunity to genotype inversions in multiple individuals, to better understand their structure and frequency. I selected eight HuRef regions from 1.1 to 21.9 kb and explore the characteristics of these loci across human and primate populations. Interestingly, I found that the structures of submicroscopic inversions could be complex, and were often accompanied by gains and losses of DNA.

Finally in Chapter V, I describe some of the remaining technical challenges in discovery and characterization of structural variation in sequencing studies. I also explore some upcoming

technologies and their characteristics. I discuss some on-going population-based sequencing initiatives and how they can improve our understanding of genomic variants.

CHAPTER II: TOWARDS A COMPREHENSIVE STRUCTURAL VARIATION MAP OF AN INDIVIDUAL HUMAN GENOME

Data from this chapter have been included in the following publication:

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, Kirkness EF, Levy S et al. 2010. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11(5):R52.

I performed the mate-pair, split-read, and Agilent 24M variant calling, the generation of a non-redundant variant set, data-mining of variants from personal genome studies, and cross-platform and cross-study comparison. The comparison with genomic features was done by Jeff MacDonald. The variant-calling of the NimbleGen 42M array variant-calling was done by myself, Drs Dalila Pinto, Donald Conrad and Matthew Hurles, while the Agilent 24M by myself, Drs Hansoo Park and Charles Lee. The Sanger sequence alignment was done by Dr. John Wei. PCR and qPCR validation experiments were performed by me and Dr. Muhammad Rafiq.

II.A Introduction

Comprehensive catalogues of genetic variation are crucial for genotype and phenotype correlation studies, in particular when rare or multiple genetic variants underlie traits or disease susceptibility (Bodmer and Bonilla, 2008; Feuk, et al., 2006a). Since the publication of the first personal genome, the HuRef genome, in 2007, several personal genomes have been sequenced, capturing different extents of their genetic variation content (Table I. 2). Comparing with HuRef, other individual genome sequencing projects identified similar numbers of SNPs, but significantly fewer structural variants (ranging from ~1,000 to ~400,000). It is clear that even with deep sequence coverage, annotation of structural variation remains very challenging, and the full extent of structural variation in the human genome is still unknown.

Microarrays (Iafrate, et al., 2004; Redon, et al., 2006; Sebat, et al., 2004) and sequencing (Khaja, et al., 2006; Kidd, et al., 2008; Korbel, et al., 2007; Tuzun, et al., 2005) have revealed that structural variation contributes significantly to the complement of human variation, often having unique population (Conrad, et al., 2010b) and disease (Buchanan and Scherer, 2008) characteristics. Despite this, there is limited overlap in independent studies of the same DNA source (Harismendy, et al., 2009; Scherer, et al., 2007), indicating that each platform detects only a fraction of the existing variation, and that many structural variants remain to be found. In a recent study using high-resolution CGH arrays, the authors found that approximately 0.7% of the genome was variable in copy number in each hybridization of two samples (Conrad, et al., 2010b). Yet, these experiments were limited to detection of unbalanced variation larger than 500bp, and the total amount of variation between two genomes would therefore be expected to exceed 0.7%.

My objective in the present study was to annotate the full spectrum of genetic variation in a single genome. The assembly comparison method presented in the initial sequencing of this genome (Levy, et al., 2007) discovered an unprecedented number of structural variants in a single genome; however, the approach relied on an adequate diploid assembly. As there are known limitations in assembling alternate alleles for structural variation (Levy, et al., 2007), for example, 44 – 52 % of heterozygous indels were estimated to have been missed due to low sequence coverage. I expected that there was still variation to be found. In an attempt to

capture the full spectrum of variation in a human genome, this current study uses multiple sequencing- and microarray-based strategies to complement the results of the assembly comparison approach in the Levy et al. (Levy, et al., 2007) study. First, I detect genetic variation from the original Sanger sequence reads by direct alignment to NCBI Build-36 assembly, bypassing the assembly step. Furthermore, using custom high density microarrays, I probe the HuRef genome to identify variants in regions where sequencing-based approaches may have difficulties (Figure II. 1). I discover thousands of new structural variants, but also find biases in each method's ability to detect variants. My collective data reveals a continuous size distribution of genetic variants (Figure II. 2a) with ~1.58% of the HuRef haploid genome encompassed by structural variants (39,520,431bp or 1.28% as unbalanced structural variants and 9,257,035bp or 0.30% as inversions) and 0.1% as SNPs (Table II. 3, Figure II. 2). While there is still room for improvement, my result gives the best estimate to date of the variation content in a human genome, provides an important resource of structural variants for other personal genome studies and highlights the importance of using multiple strategies for structural variation discovery.

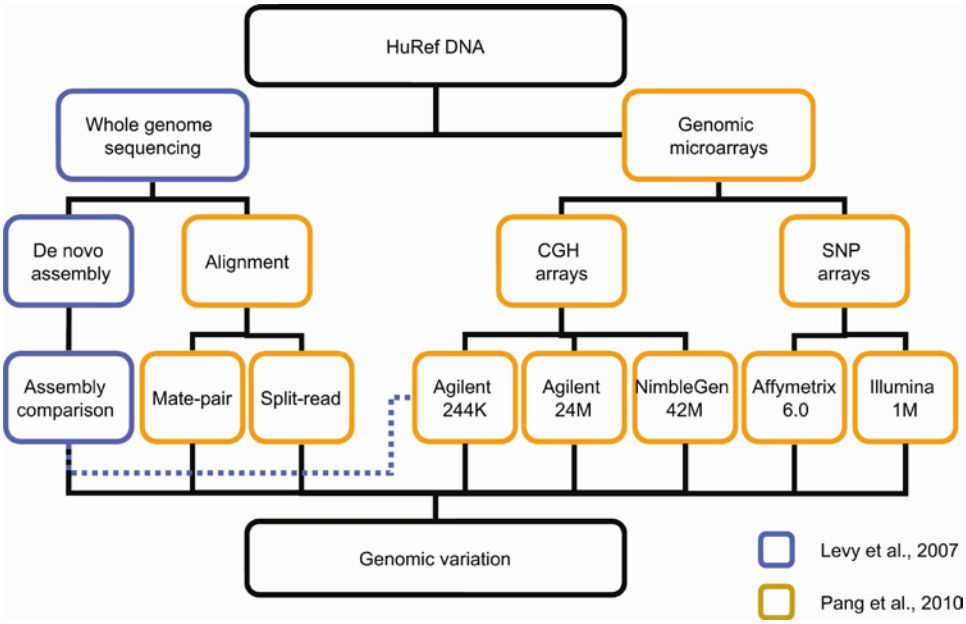


Figure II. 1. Overall workflow of the current study.
Two distinct technologies were used to identify structural variation in the HuRef genome: whole genome sequencing and genomic microarrays. The sequencing experiments, the construction of the HuRef genome assembly, and the assembly comparison with NCBI Build-36 (B36) reference had been completed in previous studies (Khaja, et al., 2006; Levy, et al., 2007; Venter, et al., 2001). Hence, these experiments are shown as blue boxes. The scope of the current study is denoted in orange boxes. I re-analyzed the initial sequencing data, and searched for structural variants in sequence alignments by the mate-pair and split-read approaches. Also, three distinct CGH array platforms were used: Agilent 24M, NimbleGen 42M and Agilent 244K. Unlike the other array platforms, which were designed based on the B36 assembly, the Agilent 244K targeted scaffold segments unique to the Celera/Venter assembly. To denote this, this figure shows a dotted line connecting between the assembly comparison outcome and the Agilent 244K box. Finally, the Affymetrix 6.0 and Illumina 1M SNP arrays were also used in the present study.

II.B Material and methods

II.B.1 Sequencing based analysis

The sequence data of the HuRef genome used for analysis was originally produced through experiments performed in the Venter et al. and Levy et al. studies (Levy, et al., 2007; Venter, et al., 2001). The sequence trace data and information files were downloaded from NCBI. In this study, I aligned 31,546,016 HuRef sequences to the NCBI human genome assembly Build-36 using BLAT (Kent, 2002). For paired-end mapping, the optimal placement of clone ends was determined by a modified version of the scoring scheme used in Tuzun et al. (Tuzun, et al., 2005). I categorized mate-pairs that mapped less than three standard deviations from the expected clone size as putative insertions, greater than three standard deviations as putative deletions, and in the wrong orientation as putative inversions. I required each variant to be confirmed by at least two clones, and for indels, I required the clones to be from libraries of the same average insert size (2kb, 10kb or 37kb) (Table II. 1). To identify small variants, the read alignment profiles were further examined for an intra-alignment gap with size greater than 10bp. Two independent “split-reads” were required to call a putative variant.

Table II. 1. Clone library information.

Library category	Average - 3 stdev (bp)	Average size (bp)	Standard deviations (bp)	Average + 3 stdev (bp)	# clones
2 kb	1,592.00	1,925.00	111.00	2,258	3,394,197
10 kb	7,936.00	10,201.00	755.00	12,466	1,887,943
10 kb	7,789.71	11,659.08	1,289.79	15,528.45	1,214,872
10 kb	7,212.37	9,290.53	692.72	11,368.69	1,615,238
10 kb	10,743.45	16,095.42	1,783.99	21,447.39	328,128
10 kb	8,454.40	12,025.12	1,190.24	15,595.84	790,319
10 kb	6,677.29	9,672.07	998.26	12,666.85	661,344
37 kb	28,555.43	37,506.11	2,983.56	46,456.79	375
37 kb	25,543.59	37,323.66	3,926.69	49,103.73	767
37 kb	25,445.28	36,476.82	3,677.18	47,508.36	48,940
37 kb	26,529.49	36,264.40	3,244.97	45,999.31	208,810
37 kb	26,724.88	35,581.33	2,952.15	44,437.78	321,221
37 kb	29,518.61	39,459.32	3,313.57	49,400.03	360,879
37 kb	25,503.63	36,579.63	3,692.00	47,655.63	43,337
37 kb	26,082.21	36,835.83	3,584.54	47,589.45	212,222
37 kb	25,654.31	36,427.13	3,590.94	47,199.95	63,152
37 kb	26,421.54	37,346.61	3,641.69	48,271.68	65,018
37 kb	26,404.97	35,517.74	3,037.59	44,630.51	33,687
37 kb	27,590.97	38,557.74	3,655.59	49,524.51	84,879
37 kb	26,062.32	37,265.88	3,734.52	48,469.44	1,906
37 kb	27,811.23	38,209.23	3,466.00	48,607.23	1,920
37 kb	25,775.42	36,640.55	3,621.71	47,505.68	1,882
37 kb	25,478.17	36,354.73	3,625.52	47,231.29	1,917
37 kb	26,940.34	37,452.61	3,504.09	47,964.88	1,918
37 kb	26,988.90	37,293.48	3,434.86	47,598.06	1,152
37 kb	26,307.07	38,392.00	4,028.31	50,476.93	384
37 kb	26,055.90	38,360.31	4,101.47	50,664.72	383

II.B.2 Array based analysis

An Agilent 24 million features CGH array set (Agilent 24M) was designed with 23.5 million 60-mer oligonucleotide probes tiled along the NCBI Build-36 assembly. The HuRef genomic DNA was co-hybridized with the female sample NA15510 from the Polymorphism Discovery Resource (Scherer, et al., 2007). The statistical algorithm ADM-2 by Agilent Technologies was used to identify CNVs based on the combined $\log_2$ ratios. Similar experimental procedures and analyses are described in other studies (Kim, et al., 2009; Park, et al., 2010). Additionally, a custom NimbleGen 42 million features CGH microarray (NimbleGen 42M) was used in this study, and its design, experimental procedures and data analysis had been described in detail elsewhere (Conrad, et al., 2010b; Scherer, et al., 2007). HuRef genomic DNA was also co-hybridized with the sample NA15510. For both the Agilent 24M and NimbleGen 42M arrays, CNVs with >50% reciprocal overlap and opposite orientation of variants identified in NA15510 in Conrad et al. were removed, as these were specific to the reference.

The HuRef sample was also run on the Affymetrix SNP Array 6.0 and Illumina BeadChip 1M genotyping arrays. I followed the protocol recommended by the manufacturers. For Affymetrix 6.0, the default parameters in the BirdSeed v2 algorithm were used to perform SNP calling. Partek Genomics Suite (Partek Inc.), Genotyping Console (Affymetrix, Inc.), BirdSuite (Korn, et al., 2008) and iPattern (Zhang J et al., manuscript submitted) were used to call CNVs. For Illumina 1M, the SNP calling was done using the BeadStudio software. QuantiSNP (Colella, et al., 2007) and iPattern were used to identify CNVs. For both platforms, only variants confirmed by at least two calling algorithms were included in the final set of calls.

The Agilent Custom Human 244K CGH array (Agilent 244K) was designed to target 9,018 sequences >500bp in length that were annotated as “unmatched” sequences in Khaja et al. (Khaja, et al., 2006). CGH experiments were performed with genomic DNA from HuRef and six HapMap samples, hybridized against reference NA10851. Feature extraction and normalization were performed using the Agilent feature extraction software. The programs ADM-1 in the DNA Analytics 4.0 suite (Agilent Technologies), and GADA (Pique-Regi, et

al., 2008) were independently used to call CNVs, and those that were confirmed by both algorithms were then used in this study.

II.B.3 Non-redundant variant data set

To generate a non-redundant set of HuRef variants, I combined the lists of structural variants generated. For CNVs, to determine if two calls are the same, I required that they shared a minimum of 50% size reciprocal overlap; for inversions, I required that they shared at least one boundary. For those calls that were indicated to be the same variant, I recorded the one with the best size/boundary estimate (with preference given to assembly comparison, then split-read, NimbleGen-42M, Agilent 24M, mate-pair, Affymetrix 6.0, and Illumina 1M, in that order). For this analysis, I excluded variants called in the Custom Agilent 244K arrays.

II.B.4 Polymerase chain reaction (PCR) and quantitative real-time PCR validation

I used multiple approaches to validate structural variants found in this project. PCR primers were designed to target flanking sequences of indels detected by sequencing-based methods, such that PCR products representing the different alleles can be differentiated on a 1.5 % agarose gel. DNA from HuRef and five HapMap individuals of European ancestry were tested in PCR experiments. Amplifications and deletions detected by CGH arrays were tested by quantitative real-time PCR (qPCR). DNA from HuRef and six additional control individuals were used to assess the variability in copy number. Each assay was run in triplicate and the *FOXP2* gene was used as the reference for relative quantifications. See Table II. 2 for all primer sequences.

Table II. 2. List of validated variants and their primers and probes

Mate pair Chrom	Start	End	Size	Type	PCR/qPCR	Forward primer	Reverse primer
chr6	160,413,658	160,419,699	6,250	del	qPCR	GGAGATGTCAGGAGTGTGTAAGG	AGGACGCTACCAAAGAGCTTC
chr16	67,273,862	67,278,219	3,242	del	qPCR	TGGACAATGGTGAGAGCATC	AAATAGCCCTCCTCGACACA
chr9	27,179,804	27,186,618	3,036	del	qPCR	GATGTGGCACCAGGAGAATTA	TCTGAAACAGAATCCTCTGCAA
chr18	53,097,735	53,099,784	2,738	del	PCR	ATCCAATAAGGTGCCTCTGC	CAAGGGAGATGCCTACAGGA
chr9	130,450,741	130,454,363	2,694	del	PCR	ACGGGAGACAGACCTGAATG	AGAAGCATTGCCCTTTGAGA
chr9	109,072,830	109,075,329	2,627	del	PCR	TGCTCAGACTGCACTTCCAA	GCATTTCAGCATCCCACAAT
chr4	165,860,789	165,863,310	2,574	del	PCR	ATGGTAGGATGCCGTCATTG	CGGAGGACTGTGAATGTTGA
chr5	78,461,229	78,462,722	2,465	ins	PCR	TCCTCCCCTAGCTTTGGTCT	ATGATGACAATGGCGGTTTT
chr15	51,112,661	51,115,499	2,452	del	PCR	TGATTTTGTATCATGATCAGCTTG	AATCATTTGGGCTGGCTCT
chr4	128,132,938	128,134,943	2,433	del	PCR	TGCCAGCTTACTCACCCTCT	TTCCACCTCCCCCTTCTATT
chr3	3,206,919	3,207,649	2,393	ins	PCR	CAAATGCAAGAGGCATTTCT	TGGGGAACTAAGGCTTATTGG
chr2	239,165,137	239,166,987	2,355	del	PCR	GGAACTCAGGCTTTTCAACG	AGGAGATTCAGGCCATTCCT
chr8	11,282,662	11,284,635	2,168	del	PCR	TGCCTTTCACTTTTGCCTCT	ACCCATCCCTGCTTCTCTCT
chr4	58,581,984	58,583,707	1,100	del	PCR	GGAAATTGCTAGATTGGTGGA	CTTCCGTACAGAGCCCATTT
chr12	8,908,157	8,909,134	1,090	del	PCR	ACATATTGTCCTTGCCTTCTCG	TCTCCTGATCTCTTCAAGTCCA
chr4	139,185,478	139,186,884	1,030	del	PCR	CCCTCAGGAAGGGAGACATA	GTGCACGTGAAGCAGATTGT
chr7	157,721,887	157,723,436	962	del	PCR	GGAAACGCTTGGCTAAATGA	GGAGCAGCCAGCACAGGT
chr7	316,083	317,193	936	del	PCR	GCAAAGGATGCAGGAGGAG	CAGTGGGTTTGGGAGGTG
chr3	101,110,993	101,112,696	880	del	PCR	CCATTTCCATTCAACTGCCTAT	CCAGTCTCCATCTCTTCATCCT
chr4	190,809,808	190,811,339	750	del	PCR	AAGCGTGCAGTCATGGAAGT	CCCACACTTTCCAGCTTGTT
chr3	185,359,277	185,360,459	686	del	PCR	CTCAGCCTTTGAGGTGCTAGTT	ATACTGCAATCCAGGAAGTGGT
chr11	60,328,009	60,329,299	622	del	PCR	GGAGAAAGAGCACCTGGAACTA	CTGCAAACAGAGTGGAGTGAAG
chr2	195,922,788	195,923,944	516	del	PCR	TCCACTTCCAACTCTAGCATGA	CGATCAACCAAGCATATAACCA
chr6	137,354,930	137,356,192	516	del	PCR	AAGTTGAAGGTGCCCACTGA	GCCATAGACATGCCCTTTGT
chr12	45,361,232	45,362,078	498	del	PCR	TAGCCAATTTCATTGCCTTGT	CAAATGTTCAAAATCTGGGTACG
chr18	66,997,320	66,998,367	486	del	PCR	TCCTCTTCCTAAGGCATATTTCA	GGAAAAGTCTGAGGCTGCAT
chr6	46,075,421	46,076,160	462	del	PCR	TGGGTATTTCAGTTTTGGACCT	AATCAGATGGATACTCCCCTCA
chr3	96,598,848	96,599,246	461	del	PCR	GATGCGAAAACACCTGCATA	TGAGCAATGGCCTACAGAAA
chr4	43,446,460	43,447,145	446	del	PCR	GTGGGCTGCTTCTGAACATT	CTTCCTGTGTCCAGCCTGAT
chr1	196,160,273	196,161,167	437	del	PCR	GCCCTCTAGGGATGAGAAGAA	CCCACTCCACACCAAGGTAA
chr12	27,324,607	27,325,754	424	del	PCR	ACTGGCTGCCATTCAACG	CACCACTGTTGTTGTCATGGA
chr5	11,391,901	11,392,525	422	del	PCR	GCATGATTGCCTGTCTCTCA	TGCCTTATTTCCCCATAGCA

chr10	46,508,167	46,508,929	409	del	PCR	CAGCTTGAGCTGAAAGATGC	GGATGGAGTAGGAAGTGAGT
chr10	132,048,184	132,048,985	400	del	PCR	CTGAATCCAAGGCTGTGCTT	TCCCAACCAACCAAACTACC
chr18	35,315,498	35,316,327	395	del	PCR	CACTTCCTCAGCGTGTGCTA	GATTCTCAGCCTGGTGAAGG
chr12	121,576,220	121,576,983	388	del	PCR	TCCAATCGTTAGGAAGAGCAA	TTTGTACGCTCAATGAGATGCT
chr4	133,333,257	133,333,943	384	del	PCR	AGAATGCCCAGGACTGACTT	TTACTAGGCGAGTTCCCCATT
chr13	21,952,874	21,953,674	383	del	PCR	GGTGTTTTCTTCTATATGGACAATCA	TGTAGTAACTGGAACTAAGTGAACACA
chr6	169,970,843	169,971,587	374	del	PCR	GCTGGCCACAAGTATGCAG	CCTGGCCACAATTCTCCAA
chr7	103,974,488	103,975,247	370	del	PCR	TGCATAACTGAGCTGGGAGA	AATTGACTGCAACCTCAAGGA

Split read

Chrom	Start	End	Size	Type	PCR/qPCR	Forward primer	Reverse primer
chr4	139,185,661	139,186,667	1,007	del	PCR	GTATGCCTGGATTCTTTCAAGTG	GGTTGGCTGTAAGTTTGGTAGTG
chr6	109,535,615	109,536,423	811	del	PCR	CAGGTTTAGAACTCAAGAATTTGG	AAATCACAGCACAAGGTCTCA
chr9	103,483,792	103,484,557	766	del	PCR	CCTTTGTCTGTGAATTTGTTCC	GTGTTGAACTATGGTCGGGTAG
chr17	27,633,422	27,634,030	609	del	PCR	ATTTTGTTCCTGGGCATCAG	TGGAGATGCAGCTGTGAGTT
chr3	121,376,675	121,377,134	460	del	PCR	GCTCTCAGCTCACCATCCTACT	CCTGCTGGACCCTGTTAAAGA
chr5	2,004,012	2,004,454	443	del	PCR	GCGTAGTCTCTACCCTCACACC	GCAGACAATGGACAATGCAC
chr2	216,866,591	216,866,971	381	del	PCR	TGCTTCCGTCTTCATGGAAT	GGAGAACTGAGAGCAGGTCAG
chr7	28,180,827	28,181,190	364	del	PCR	TGTCACCTACGGGAAGAATTG	GACTGGGAAAGTGGTGTCAAA
chr13	81,787,199	81,787,541	343	del	PCR	CCGTACCGTATGACTAAGAACCA	CGCAAGACACAGGCTATCATTA
chr3	193,338,785	193,339,122	338	del	PCR	TCCAGGTCTTGGTTGACTTACA	TATACCACTGAGGAGGCAACAA
chr4	36,146,313	36,146,645	333	del	PCR	AACAATTTGGTCTTTGGTCCTG	TCTCTGGCCTTAGTGTCAGCTT
chr1	230,489,745	230,490,072	328	del	PCR	CAAGCTGGTGATCTCTCACTGT	CAAGTGCAGCCTGTCCTCTT
chr7	30,502,113	30,502,440	328	del	PCR	CACTGCATTCCACAGAGACATT	TTGTAGCTCACTTGATGATGGTG
chr5	122,302,244	122,302,566	323	del	PCR	TTCCAGGATTCTCTCTTCTCTCAG	GTTGATGTGGCAGCAGCAGT
chr2	78,818,917	78,818,917	315	ins	PCR	CTGCAAGGGAAAGATGAGCTA	CATCCAAACATAACATCCATACAAG
chr5	16,769,612	16,769,612	310	ins	PCR	GAACATTACCGAGCCTTCCATA	ACGTCTTCCTGGGCTCTTCT
chr3	139,048,074	139,048,074	306	ins	PCR	ATCTCCACCCTCAAGCTCTTC	AGGGCAGAAATACTCAGTCCAG
chr9	89,672,635	89,672,635	289	ins	PCR	TGCATCGACACTGAGAAATACA	CTGTGATGACTTTGGAAGCACT
chr12	1,515,635	1,515,635	262	ins	PCR	GGTATATGCTGGGACGAGGA	CCTTTTTAAGAATAACAAAATGCAA
chr18	10,452,407	10,452,654	251	del	PCR	TTGCTAAACACTGGGATGATTG	GGAACAGCTCCCAACAAACTAC
chr10	63,413,623	63,413,800	178	del	PCR	CTGACCTGCGTATGAAGCACT	AAACGTAAAGCACATCCCAGAG
chr5	5,984,072	5,984,072	148	ins	PCR	CAGGTGGTCAGGAATCATGTG	GTGAGTTCTGGAATGGAGTCCT
chr18	72,923,639	72,923,639	144	ins	PCR	ATAGCCAGGACCCAGTGAGAT	CTCTATCCCGAAGCAGAATAGC
chr3	52,841,692	52,841,692	135	ins	PCR	GGAGGAGGCAGTGTGAACAA	GTCTCAGGGACCACCTCTCCT

chr17	78,490,010	78,490,010	124	ins	PCR	GCAGTGAGACATGATGACCAG	GATCCCAGCAGTGAAGATGAC
chr11	80,825,778	80,825,891	114	del	PCR	GCTTGTCTCCTTGCTCAACAG	GTGCCATTGCTCTGAATGACTA
chr17	46,020,716	46,020,718	111	ins	PCR	ACACGTTGTGTATCCAGTCCAC	AACATGGACCGCACAGGTAT
chr20	62,179,973	62,179,973	104	ins	PCR	AGAGACAGTGCTGGAGTTCTGC	TCGGTCTCCTCATCTACTTGCT
chr9	72,273,848	72,273,848	101	ins	PCR	TGTGTTGGCACCCATTATTTAC	CACAGGCCCATTAGATGTAGC
chr16	87,005,900	87,005,900	99	ins	PCR	GAAGAGACGAAACCCTGATGTT	GAGACAGAGGCGAGACCAAC
chr17	69,685,307	69,685,401	95	del	PCR	GACAGTGTGTGCCTGCATGT	GACGGACACAACAGCATACG
chr2	240,757,503	240,757,503	84	ins	PCR	CAGCAGAAGCTGACCCTGT	GGTCATTGCTGTGCCCTACT
chr20	62,322,579	62,322,659	81	del	PCR	AACTGTGGTCTGGGAAGCAG	CGGATAATTGGACCTCGTTG
chr15	99,543,140	99,543,140	80	ins	PCR	GTGTGGAATGGGAGATGGAG	GGCAAAGATTCCATCCAGAA
chrX	110,008,550	110,008,550	75	ins	PCR	AGGGATTTGCTGGTCAATGT	TGCTCACTCAATTCTCATTTCC
chr21	17,379,379	17,379,451	73	del	PCR	CCCTGAGTATTAAAGGCAAATCC	AATAACAACACCAGTGACCCAAG
chr18	61,768,283	61,768,283	72	ins	PCR	TGAGCTTACAAATTGCCACAA	CACCAAAGAAAGAGATGACTTGA
chr13	46,191,128	46,191,128	70	ins	PCR	CCAAGGAGAAGTTGAGAGCTACA	GCTCAGTCTGTTTGGAGAACG
chr1	244,723,725	244,723,789	65	del	PCR	CCAATGACTTCACAGAGCAGTAG	ACGTTGTTGACCGATTAAAGG
chr21	46,355,933	46,355,994	62	del	PCR	CCCGCTTCCTTGAAGACTG	TCAGGAGGGACTGCTGTTG
chr16	61,621,368	61,621,368	56	ins	PCR	TAAACCCTATGAACGCTGAGG	ACTATAAGGCCGGACAGAAGAA
chrX	66,681,884	66,681,928	45	del	PCR	GGATGGAAGTGCAGTTAGGG	GCTGCTGTTGCTGAAGGAGT
chr17	18,795,453	18,795,453	39	ins	PCR	ATCTGCACTGAATCCTACTCTGC	CAATATCTGTCACACACCAAGGA
chr14	73,332,432	73,332,432	30	ins	PCR	GTCCATAAGGGGCTGCAGAT	GCTCTTTGCAACCAACTCAG
chr11	22,171,452	22,171,458	24	ins	PCR	GCCTGGAGAAGTACTGGGAGA	CGCCAACACTTCCAGGAGAT
chr5	79,986,487	79,986,487	18	ins	PCR	AAGCCTGAAATCCACCTCCT	CTTCCCACCTTCCCCTTCT
chr11	6,368,510	6,368,510	13	ins	PCR	CCGAGAGATCAGCTGTCAGA	TGATGGCGGTGAATAGACCT
chr6	16,435,893	16,435,893	12	ins	PCR	CCTCCCGAGGGACAAAGT	CGTGCAGTACGCTCACCTG
chr15	88,121,138	88,121,149	12	del	PCR	AGAGCCTGACCAAGATCGAG	AAGGGCCTCCACACATACCT

Agilent custom 244k								
Celera_scaffold	**Start**	**End**	**Size**	**Type**	**PCR/qPCR**	**Forward primer**	**Reverse primer**	**PCR size**
GA_x5YUV32V2AG	9,803	18,539	8,737	dup	qPCR	TGCACTGTAGTAAGGCAAGGAC	TTGATGCCAAGTAGTATTGAGTGTC	98
GA_x5YUV32VQPK	7,692	14,813	7,122	dup	qPCR	TTGGAGCTGAGAACACAGGTAA	TGGTAGTCAGTCAAGGTCATGG	123
GA_x5YUV32W46Y	33,940,348	33,947,171	6,824	dup	qPCR	CAGAGGCTATCTGCTCCTTGTA	TGAATGCAGGAGTGATGAAGAG	112
GA_x5YUV32VTY6	13,803,851	13,808,084	4,234	dup	qPCR	AATGAAGCCAGAGTATGCAAGG	TCAAGGTTAGCTCATGGTCACT	125
GA_x5YUV32VS0N	371	3,047	2,677	dup	qPCR	GAACTCAACTCTCTGCACCTGA	CTCTGTCACTCCAGCCTCATCT	135
GA_x5YUV32V15N	381	2,638	2,258	del	qPCR	GAGGCAGAAGAACAGGACAGAC	CTGAGAGTCCAGAAGCAGCTC	129
GA_x5YUV32W3JR	1,511,235	1,513,331	2,097	dup	qPCR	TGGTGGGTCTTGGACTCTTACT	AGAACTCAGAGGCAGCTTTCAT	125

GA_x5YUV32V1HL	307	1,654	1,348	dup	qPCR	GGCGAGTATGTATGTGCAGTTG	ATTGGTCAGTATCCATCAACATT	112
GA_x5YUV32V00B	2,293	3,591	1,299	dup	qPCR	GCTCACTGGCCTGATTACTGAT	CAACTTCAGAACTGGTAGAGAATG	119
GA_x5YUV32UWVA	15,545	16,821	1,277	dup	qPCR	CTGAGGTTCAGCGAGGTTCT	AAGTCACTGCTACCACGAAGGT	114
GA_x5YUV32W7K2	46,804,628	46,805,061	434	del	qPCR	TCACCTCGATCTAACATCTGGA	GTCACTGAATTGCATCGTGATT	99

NimbleGen 42M

Chrom	Start	End	Size	Type	PCR/qPCR	Forward primer	Reverse primer
chr3	163,994,833	164,109,038	114,206	dup	qPCR	ATTCCCAGGTCTTAGCCTTCTC	TAAGCCTTTCATCTTCCTTCCA
chr13	56,651,124	56,686,891	35,768	del	qPCR	ACTTTATGGGCAGTAGCACGAC	GCCAGGCAATTAAATCTCACTC
chr2	88,913,483	88,942,451	28,969	dup	qPCR	CCAATGTCCAGGCATCATTC	AGAGACAGCAGCTTGGCATACT
chr4	9,801,723	9,814,213	12,491	del	qPCR	AAGTGAGGGCTCCATCTCATAA	CTAAGATCGCTGACAATGATTCC
chr15	22,971,787	22,981,757	9,971	del	qPCR	AAGGGCTTCCTTCAACTCAAT	GACAGAGAGCACCCTCATAACA
chr17	49,208,550	49,214,575	6,026	del	qPCR	AATATCAGCCTAGTTTGTCTTCCAG	GGATGGTCAGTAATCCATACACAA
chr2	178,552,127	178,556,191	4,065	del	qPCR	ATGGAACTTAATGCCCAAACAC	TTTAGGTTGTACCCATGATTGC

Agilent 24M

Chrom	Start	End	Size	Type	PCR/qPCR	Forward primer	Reverse primer
chr8	39,351,157	39,505,456	154,300	del	qPCR	CCACTGGACACTCACAGCTT	TGTGCCTGGCTAACACATTC
chr3	163,995,557	164,109,021	113,465	dup	qPCR	ATTCCCAGGTCTTAGCCTTCTC	TAAGCCTTTCATCTTCCTTCCA

Mate pair

Chrom	Start	End	Size	Type	FISH	Probes
chr16	21,501,943	22,620,002	1,118,060	inv	FISH	G248P81317B4 G248P89030H10 G248P85584B2 G248P8661A9

II.B.5 FISH validation

To validate large variation, FISH experiments were performed using fosmid clones as probes on lymphoblastoid cell line from HuRef and seven other HapMap individuals. Five metaphases were first imaged to check for correct chromosome localization and hybridization, and then interphase FISH was performed to validate predicted inversions, similar to the protocol outlined in the Feuk et al. study (Feuk, et al., 2005) with the addition of the aqua probe, DEAC-5-dUTP (Perkin Elmer; NEL455).

II.B.6 Overlap analysis

Overlap with other datasets, genomic features and between subsets of data in the current paper was performed using custom PERL scripts. When comparing variants, two sites were considered overlapping if the reciprocal overlap among their estimated sizes was ≥ 50%. Data sources used for the annotations of overlaps with genomic features are listed in Appendix Table 1. To evaluate significance, 1,000 randomized sets of simulated variant calls were created and overlap analysis was performed against the same data source. For each simulation, I recorded the number of instances where I observed a higher number of overlaps than the real variant data set. A p-value was computed as the fraction of simulations whose number of overlap was greater than the number of real overlaps.

II.B.7 Structural variation imputation

Using a cutoff of 50% reciprocal overlap, there were 405 sites of overlap between the HuRef and genotyped, validated Genome Structural Variation (GSV) loci (Conrad, et al., 2010b). The best r^2 value was computed between each of those GSV CNVs and an European CEU HapMap SNP in the neighboring genomic region. Here, I defined a minimum threshold of $r^2 = 0.8$, below which the HuRef structural variants were deemed not well imputed by SNP. Detailed description on genotyping, phasing, and tagging calls onto haplotypes defined by HapMap SNPs is presented in the Conrad et al. study (Conrad, et al., 2010b).

II.C Results

Several different analytical and experimental strategies were employed to exhaustively analyze the HuRef genome for structural variation. An overview of the different analyses performed is shown in Figure II. 1.

II.C.1 Sequencing-based variation

I used computational strategies to extract additional structural variation information from the existing Sanger-based sequencing data generated as mate-pair reads from clone libraries of defined size (Levy, et al., 2007). First, I adopted a mate-pair mapping approach (Kidd, et al., 2008; Korbel, et al., 2007; Tuzun, et al., 2005) and aligned 11,346,790 mate-pairs from libraries with expected clone sizes of 2, 10 or 37 kb (Table II. 1) to the NCBI Build-36 assembly. I found that 97.3% of mate-pairs had the expected mapping distance and orientation. Mate-pairs discordant in orientation or mapping distance were used to identify variants, and I required each event to be supported by at least two clones. In total, this strategy was used to identify 780 insertions, 1,494 deletions and 105 inversions (Figure II. 1, Table II. 3 and Appendix Table 2). In an independent analysis of the same underlying sequencing data, I then captured structural variants by examining the alignment profiles of 31,546,016 paired- and unpaired- reads to search for intra-alignment gaps (Mills, et al., 2006). The presence of an intra-alignment gap in the sequence read or in the reference genome would indicate a putative insertion or deletion event, respectively. The identification of such 'split-read' alignment signature complements the mate-pair approach, as significantly smaller insertions and deletions can be discovered. I required at least two overlapping split-reads having an alignment gap >10bp to call a variant. While the initial assembly comparison study has acknowledged that due to insufficient coverage at 7.5 fold, between 44 % to 52 % of the heterozygous indels have been missed (Levy, et al., 2007), the depletion of heterozygous indels above 10 bp is especially notable. According to variant size distribution, one would expect ~ 10 % of indels to be over 10 bp (Levy, et al., 2007). Indeed, 10.4 % of homozygous indels are over 10 bp; however, surprisingly less than 3.3 % of heterozygous indels are of that size. In order to complement the assembly comparison indels, I focused my effort in finding additional calls > 10 bp – especially the heterozygous ones – by the split-

read method. A total of 8,511 insertions and 11,659 deletions ranging from 11 to 111,714 bp in size were identified (Figure II. 1, Table II. 3 and Appendix Table 3).

Table II. 3. Structural variants detected by different methods.

Method	Type	#	Min size (bp)	Median size (bp)	Max size (bp)	Total size (bp)
Assembly comparison[a]	*Homo. insertion*	*275,512*	*1*	*2*	*82,711*	*3,117,039*
	Homo. deletion	*283,961*	*1*	*2*	*18,484*	*2,820,823*
	Hetero. insertion	*136,792*	*1*	*1*	*321*	*336,374*
	Hetero. deletion	*99,814*	*1*	*1*	*349*	*250,300*
	Inversion	*88*	*102*	*1,602*	*686,721*	*1,627,871*
Mate-pair	Insertion	780	346	3,588	28,344	3,880,544
	Deletion	1,494	340	3,611	1,669,696	10,531,345
	Inversion	105	368	3,121	2,026,495	8,068,541
Split-read	Insertion	8,511	11	16	414	224,022
	Deletion	11,659	11	18	111,714	1,764,522
Agilent 24M	Duplication	194	445	1,274	113,465	1,065,617
	Deletion	319	439	1,198	852,404	2,779,880
NimbleGen 42M	Duplication	366	448	4,665	836,362	11,292,451
	Deletion	358	459	2,460	359,736	3,861,282
Affymetrix 6.0	Duplication	17	8,638	42,798	640,474	2,011,557
	Deletion	21	2,280	13,145	856,671	1,978,028
Illumina 1M	Duplication	3	11,539	22,148	87,670	121,357
	Deletion	9	8,576	32,199	145,662	431,131
Custom Agilent 244k	Duplication	44	219	1,356	8,737	98,529
	Deletion	7	170	332	2,258	4,130
Non-Redundant Total[b]	**Insertion/Duplication**	**417,206**	**1**	**1**	**836,362**	**19,981,062**
	Deletion	**390,973**	**1**	**2**	**1,669,696**	**19,539,369**
	Inversion	**167**	**102**	**1,249**	**2,026,495**	**9,257,035**

[a] I used an italicized font to distinguish the results from the Levy et al. study. Moreover, from that previous study, I included all homozygous indels, heterozygous indels, indels embedded within simple, bi-allelic, and non-ambiguously mapped heterozygous mixed sequence variants, and only those inversions whose size is at most 3Mb.

[b] Complete data is presented in Appendix Tables 4 to 6. Non-redundant variation size distribution is presented in Figure II. 2A.

II.C.2 Array based variation

I used two ultra-high density custom CGH array sets and two commonly used SNP genotyping arrays to identify relative gains and losses. A significant amount of variation was detected from the two custom CGH arrays: an Agilent oligonucleotide array set with 24 million features (Agilent 24M) (Kim, et al., 2009), and a NimbleGen oligonucleotide array set containing 42 million features (NimbleGen 42M) (Conrad, et al., 2010b). The Agilent platform identified 194 duplications and 319 deletions, while the NimbleGen array set detected 366 gains and 358 losses, ranging in size from 439bp to 852kb, in HuRef (Figure II. 1, Table II. 3, Appendix Tables 7 and 8). Furthermore, the HuRef genome was scanned by the Affymetrix SNP Array 6.0 and Illumina BeadChip 1M, and the results are summarized in Table II. 3 and Appendix Tables 9 and 10.

The majority of microarrays used for CNV analyses are designed based on the NCBI assemblies. Therefore, any region where the reference exhibits the deletion allele of an indel, or sequences mapping to gaps in the assembly, will not be targeted. In previous studies (Istrail, et al., 2004; Khaja, et al., 2006), many unknown DNA segments were identified to have no, or poor alignment to the NCBI reference when compared to the Celera R27C assembly. To capture genetic variation in such potentially novel sequences, a custom Agilent 244K array was designed to target those scaffold sequences at least 500bp in length. CGH was then performed on seven HapMap individuals and detected 231 regions (101 gains and 130 losses) in 161 scaffolds to be variable (Appendix Table 11). Of these, I found 44 gains and 7 losses in 36 Celera scaffolds specific to HuRef (Figure II. 1, Table II. 3). Using paired-end mapping, as well as cross-species genome comparison with the chimpanzee, I was able to find a placement in NCBI Build-36 for 25 of 36 scaffolds that were copy number variable in HuRef. Two of the scaffolds were mapped to regions containing assembly gaps, 15 of 25 anchored scaffolds corresponded to insertion events also detected elsewhere (Kidd, et al., 2008; Tuzun, et al., 2005), and the remaining eight represent new insertion findings (Appendix Table 12)

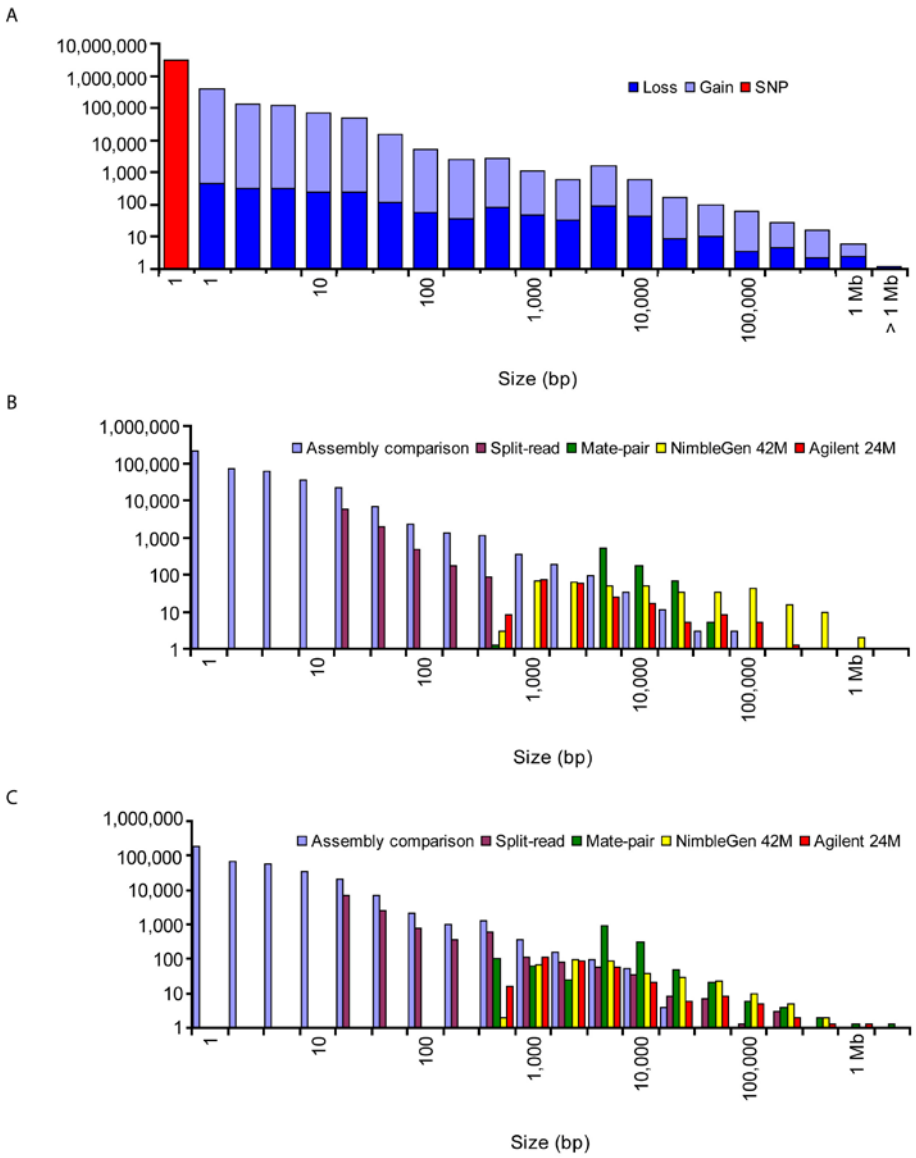


Figure II. 2. Size distribution of genetic variants.
(A) A non-redundant size spectrum of SNP and CNV (including indels) and a breakdown of the proportion of gain to loss. The indel/CNV dataset consists of variants detected by assembly comparison, mate-pair, split-read, NimbleGen 42M and Agilent 24M. The results show that the number and the size of variants are negatively correlated. Although the proportions of gains and losses are quite equal across the size spectrum, there are some deviations. Losses are more abundant at the 1 to 10kb range, and this is mainly due to the inability of the 2kb and 10kb library mate-pair clones to detect insertions larger than their clone size. The opposite is seen for large events, where duplications are more common than deletions which may be due to both biological and methodological biases. The increase in the number of events near 300bp and 6kb can be explained by Alu and L1 indels, respectively. The general peak around 10kb corresponds to the interval with the highest clone coverage. (B) Size distribution of gains (insertions and duplications) highlighting the detection range of each methodology. The split-read method is

designed to capture insertions from 11bp to the size of a Sanger-based sequence read (~1kb). There is no insertion detected in the size range between the 2kb and 10kb library using the mate-pair approach. Furthermore, large gains (≥100,000bp) cannot be identified with these present sequencing-based approaches, while these are readily identified by microarrays. (C) Size distribution of deletions.

II.C.3 Validation of findings

I used several computational and experimental approaches to validate our structural variation findings. I performed experimental validation by PCR amplification and gel-sizing and confirmed 89/96 (93%) of structural variants predicted by sequence analysis (Table II. 2). Using qPCR, 20 of 25 (80.0 %) CNVs detected by microarrays were validated, and the majority of these CNVs were from the custom Agilent 244K array covering sequences not in the NCBI assembly (Figure II. 3). In addition, one inversion was tested by fluorescence in situ hybridization (FISH) (Feuk, et al., 2005). A predicted 1.1 Mb inversion at 16p12.2 was identified to be homozygous in HuRef and in all of the 7 additional HapMap samples from four populations tested, suggesting that the reference at this locus represents a rare allele, or is incorrectly assembled (Figure II. 4). In total, 90.2 % of (110 out of 122) variants ranging in size from 12 bp to 1.1 Mb were validated, suggesting a false discovery rate of about 10 %.

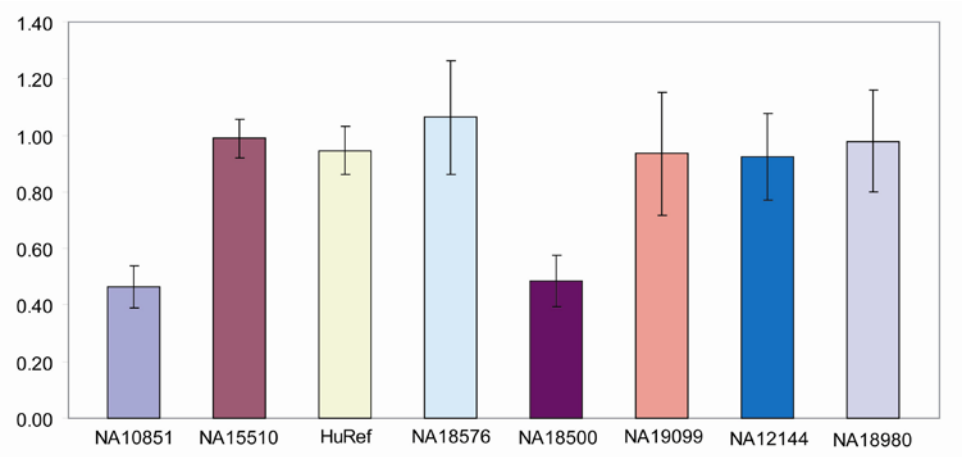


Figure II. 3. Example of a qPCR-validate gain in HuRef relative to sample NA10851 as detected by the custom Agilent 244K aCGH.
A 4.2 kb CNV was detected on the Celera scaffold GA_x5YUVVTY6, and by qPCR, I found that NA10851 had a heterozygous loss in that region, thus confirming a relative gain in HuRef. The y-axis indicates the ratio of copy number of the scaffold region versus copy number of the FOXP2 gene.

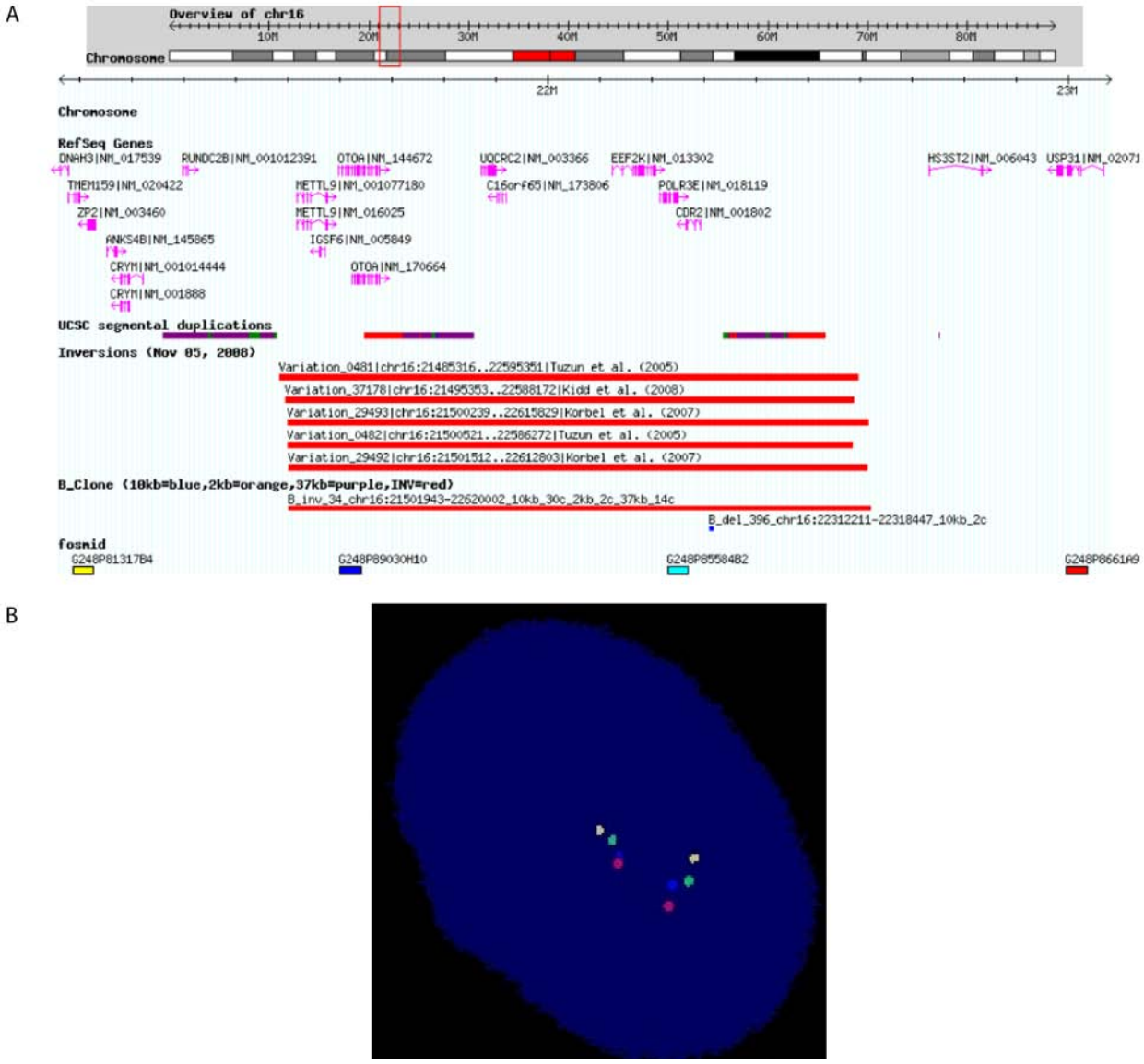


Figure II. 4. A common inversion on 16p12.2 validated by FISH.
(A) A 2Mb website schematic of the region. This 1.1 Mb inversion was detected by the mate-pair method in HuRef as seen in track "B_Clone". The track "Inversions" shows that this inversion was annotated in three other studies (Kidd, et al., 2008; Korbel, et al., 2007; Tuzun, et al., 2005). (B) An image of a four-color FISH experiment revealing that HuRef is homozygous of the 16p12.2 inverted allele. Four differentially-labeled fosmid probes were scored in >100 interphase FISH experiments and the order of the probes in HuRef were found in the vast majority of experiments (including in 7 HapMap controls from 4 different populations) to be in the yellow-green-blue-pink order. In the absence of the inversion, the order of the probes would be yellow-blue-green-pink as depicted in the assembly schematic. Therefore, as discussed in the main text this data suggests that the NCBI Build 36 reference represents a rare allele, or may be incorrect.

I then compared the structural variation identified here with the previous assembly comparison-based analysis of the same genome (Levy, et al., 2007), and found that 11,140 variants were in common. I noticed that this multi-platform method excelled in calling large variants. In fact, even after excluding all of the small variants (≤10bp) from the previous Levy et al. study (Levy, et al., 2007), I still observed that the current study tended to find larger structural variants (a current average of 1,909.3bp now versus a previous average of 113.4bp). The sensitivity of assembly comparison dropped as size increased to over 1kb, and the proportion of larger structural variants significantly increased as a result of the present study (Figure II. 5 A and B).

Finally, I determined the number of calls in this study which were either confirmed by another platform in this study, or found in the DGV (Iafrate, et al., 2004; Zhang, et al., 2006). In total, I computationally confirmed 15,642 (65.6%) of our current calls: 6,301 of which were gains; 9,726 were losses; and 65 were inversions.

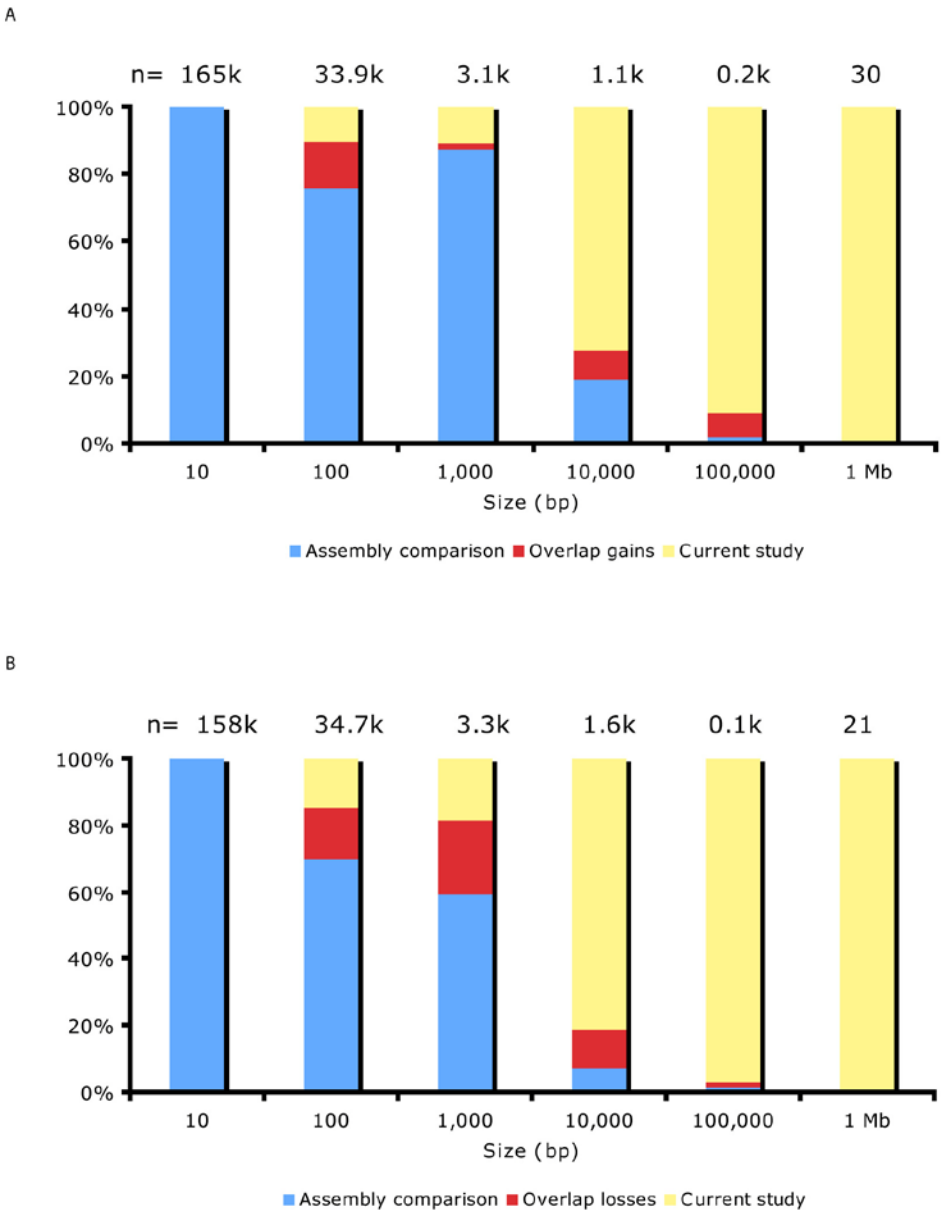


Figure II. 5. Comparative analysis of variants discovered in Levy et al. and the current study.

The two graphs illustrate the proportion of structural variants identified by the assembly comparison method, by this present combined multi-approach strategy (including mate-pair, split-read, CGH arrays and SNP arrays), and the proportion confirmed by both. The x-axis represents size range, while the numbers at the top indicate the total number of calls in a particular size range. As size increases, the number of variants called by assembly comparison decreases significantly, so this indicates that the method has limited sensitivity in detecting large calls. In contrast, current combined multi-approach strategy is more suitable in finding large variation. (A) Size distribution of gains. (B) Size distribution of losses.

II.C.4 Cross platform comparison

I performed an in-depth analysis of the characteristics of the variants detected by each of the methods. First, by contrasting against a population-based study (Conrad, et al., 2010b), I observed highly similar size estimates for the same underlying structural variants between methods (Figure II. 6). With sufficient genome coverage of clones with accurate and tight insert size, the mate-pair method yields precise variation size. Similarly, the split-read approach gives nucleotide resolution breakpoints, while the high-density CGH and SNP arrays have dense probe coverage to accurately identify the start and end points of structural variants. Overall, current multiple approaches are highly robust in estimating variant size.

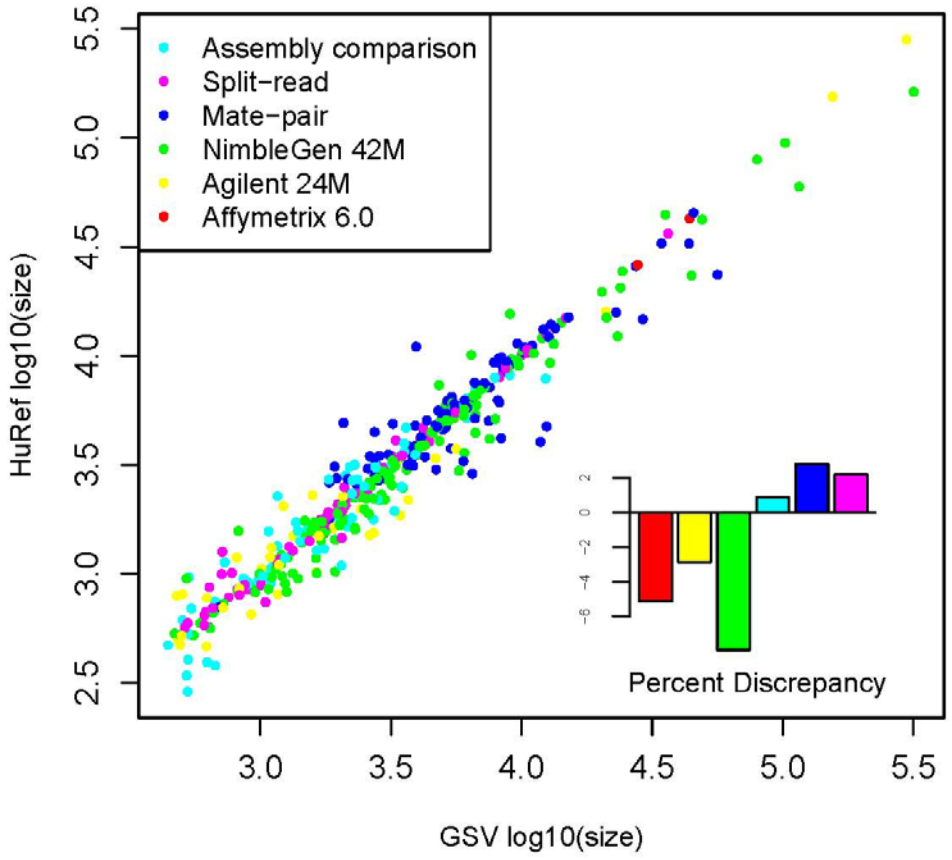


Figure II. 6. Agreement between the non-redundant set of HuRef CNVs and genotype-validated variable loci.

The agreement between sites identified by different detection methods was measured by the percentage of reciprocal overlap between the estimated size for the non-redundant set of HuRef variants and the estimated size for the CNVs generated and genotyped in the Genome Structural Variation (GSV) population genetics study (Conrad, et al., 2010b). Two sites were considered overlapping if the reciprocal overlap among their estimated sizes was ≥50%. The lower right corner plot summarizes the mean discrepancy between HuRef and GSV loci sizes, as a proportion of the GSV-estimated CNV size.

Next, I compared the variants discovered by the two whole genome CGH array sets, NimbleGen 42M and Agilent 24M, and investigated the primary reason for the discordance between the two data sets. Not surprisingly, a substantial portion of the discordant calls can be explained by the difference in probe coverage. In fact, ~70% of the unique calls on the NimbleGen 42M array had inadequate probe coverage on the Agilent 24M array to be able to call variants, and ~30% vice versa. After that, I compared the number of calls uniquely identified by the SNP-genotyping microarrays, and identified 12 and 0 novel structural variants contributed by Affymetrix 6.0 and Illumina 1M, respectively. Of the 12 new Affymetrix calls, 9 are located in complex regions containing blocks of segmental duplications.

Subsequently, when looking for enrichment of genomic features among variants detected by different approaches, I found that there was a significant enrichment ($p < 0.01$) of short SINEs in deletions called by sequencing-based approaches (mate-pair and split-read), but not in deletions called by the microarrays. Microarrays cannot detect copy number change of SINEs (e.g. Alu elements), as these regions cannot be uniquely targeted by short oligo probes, and over-saturation of probe fluorescence would prevent an accurate high copy count. Meanwhile, the sequencing methods employed here do not rely on alignments within the repeat itself, and consequently they are readily able to detecting gains and losses of these high-copy repeats. The complete result for enrichment of structural variants with various genomic features is shown in Appendix Table 1.

Finally, one of the main challenges of genome assembly is to correctly assemble both alleles in regions of structural variation. To identify heterozygous events among the split-read indels, I searched for evidence of an alternate allele. Indels were determined to be heterozygous if two or more sequence reads that would support the NCBI Build-36 allele. From the split-read dataset alone, I identified 4,476/8,511 (52.6%) insertions and 6,906/11,659 (59.2%) deletions as heterozygous. Additionally, I found that of the 10,834 split-read indels that overlapped with results from the Levy et al. study (Levy, et al., 2007), 4,332 events annotated as heterozygous in my results were previously classified as homozygous (Appendix Table 3). These differences highlight the difficulty of assembling

both alternate alleles in regions of structural variation, leading to an underestimate of the heterozygosity previously (Levy, et al., 2007).

II.C.5 The total variation content of the HuRef genome

In an attempt to estimate the total variation content in the HuRef genome, I combined the structural variants previously described in the HuRef genome in Levy et al. paper (Levy, et al., 2007) with the variants discovered in this study, to generate a non-redundant set of variants. I determined that 48,777,466bp was structurally variable, of which 19,981062bp belonged to gains, 19,539,369bp to losses, and 9,257,035bp to balanced inversions (Table II. 3). A vast majority of this variation was discovered in the current analyses (83.3% or 40,625,059bp) of the HuRef genome. Therefore, my significant contribution in detecting novel calls underscores the importance of using multiple analysis strategies for detecting structural variation in the human genome. See Figure II. 7 for the location of structural variants >1kb, and see Appendix Tables 4 to 6 for a complete list of variation in the HuRef genome.

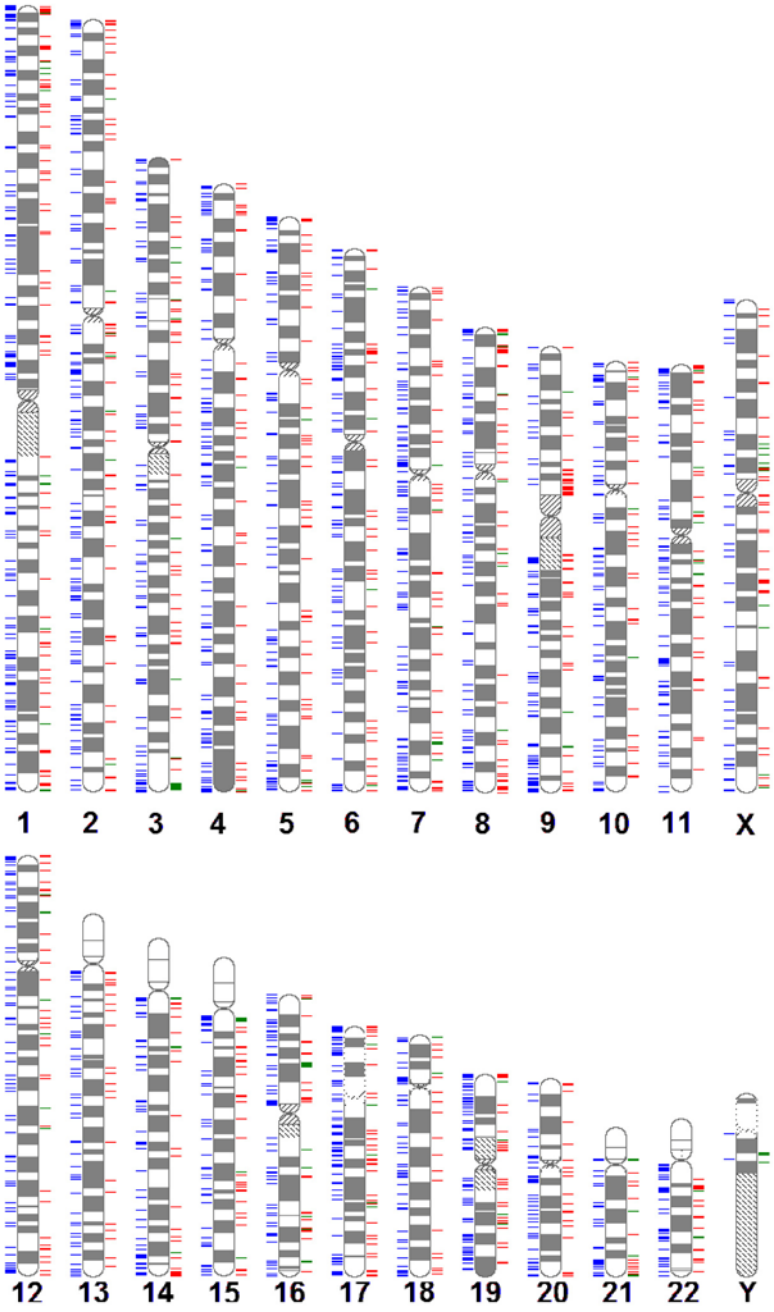


Figure II. 7. Genome-wide distribution of large structural variants in HuRef.
The sites of 2,772 structural variants whose position spans >1kb are shown. Red bars represent insertion or duplication, blue bars represent deletions, and green bars represent inversions.

II.C.6 Comparison with other personal genomes

When I compared the complete set of HuRef's structural variants with those from other published genomes (Ahn, et al., 2009; Bentley, et al., 2008; Kim, et al., 2009; McKernan, et al., 2009; Wang, et al., 2008; Wheeler, et al., 2008), I found that 209,493/808,345 (25.9%) of the HuRef variants overlapped variants described in one or more of the other six studies. Upon examining the size distribution of variants from different studies, particularly the size of insertions and duplications, I found that studies based primarily on NGS data for variation calling were unable to identify calls in certain size ranges (Figure II. 8). These results further signify that at present, NGS has notable shortcomings in structural variation detection, so additional strategies are needed to capture variants across the entire size spectrum.

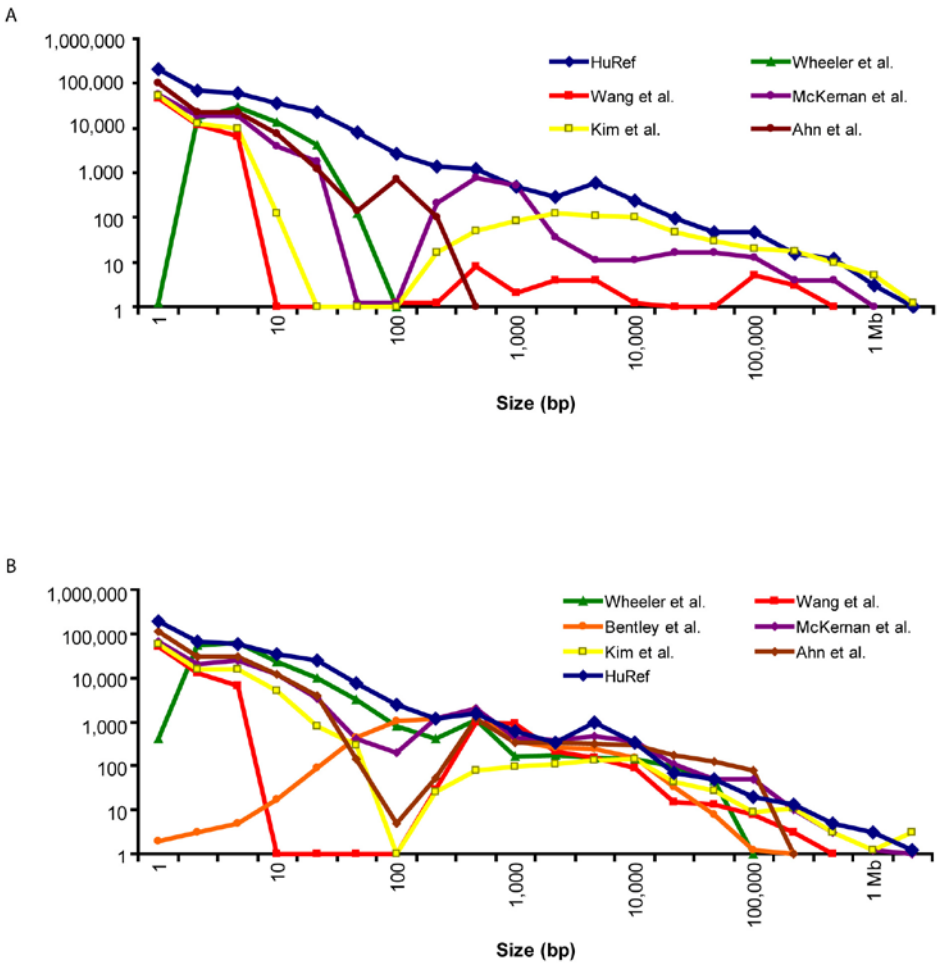


Figure II. 8. Difference in the size distributions of reported indels/CNVs in some published personal genome sequencing studies.
The graphs show variation found in a few personal genome sequencing studies. These diagrams indicate that multiple approaches are needed for better detection of structural variation. Here, the total variant set in the HuRef genome found in both the Levy et al. (Levy, et al., 2007) and the current study is displayed. Unlike the current study where the size of mate-pair indels is equal to the difference between the mapping distance and the expected insert size, the structural variants in the Ahn et al. (Ahn, et al., 2009) study is only based on the mapping distance. Besides the NGS data, I have also included the variants detected by the high density Agilent 24M data in the Kim et al. (Kim, et al., 2009) study. In Wheeler et al. (Wheeler, et al., 2008), insertions identified by intra-read alignment would be limited by the size of the sequencing read; hence, large insertions beyond the read length were not detected. Wang et al. (Wang, et al., 2008), Kim et al., and McKernan et al. (McKernan, et al., 2009) detected small variants based on split-reads and large ones based on mate-pairs and microarrays, but failed to detect variation between these size ranges. (A) Insertion and duplication size distribution. (B) Deletion size distribution.

II.C.7 Functional importance of structural variation

Next, I analyzed the complete set of structural variants in HuRef for overlap with features of the genome with known functional significance, which might influence health outcomes (Table II. 4). I found 189 genes to be completely encompassed by gains or losses, 4,867 non-redundant genes (3,126 impacted by gains and 3,025 by losses) whose exons were impacted, and 573 of these to be in the Online Mendelian Inheritance in Man (OMIM) Disease database (Appendix Tables 13 to 17). However, there was an overall paucity of structural variation ($p \geq 0.999$) overlapping exonic sequences of genes associated with autosomal dominant/recessive diseases, cancer disease, imprinted and dosage-sensitive genes. In general, there was a depletion of variation in both exonic and regulatory sequences, such as enhancers, promoters and CpG islands, in the genome of this individual.

Table II. 4. Genomic landscape and structural variants in the HuRef genome

	Total Non Redundant gains[b]			Total Non Redundant losses[c]		
Genomic Feature (# entries)[a]	**# (%) Genomic Features**	**# (%) Structural Variants**	**P-Values**	**# (%) Genomic Features**	**# (%) Structural Variants**	**P-Values**
RefSeq Gene Loci[d] (20,174)	14,268 (70.72%)	159,250 (38.17%)	0.000	13,951 (69.15%)	149,568 (38.26%)	0.000
RefSeq Gene Entire Transcript Loci[e] (20,174)	101 (0.50%)	41 (0.01%)	0.000	91 (0.45%)	47 (0.01%)	0.000
RefSeq Gene Exons[f] (20,174)	3,126 (15.50%)	3,890 (0.93%)	0.999	3,025 (14.99%)	3,723 (0.95%)	0.999
Enhancer Elements (837)	80 (9.56%)	85 (0.02%)	0.999	84 (10.04%)	93 (0.02%)	0.999
Promoters (20,174)	2,007 (9.95%)	2,071 (0.50%)	0.999	1,812 (8.98%)	1,922 (0.49%)	0.999
Stop Codons[g] (30,885)	225 (0.73%)	99 (0.02%)	0.000	272 (0.88%)	134 (0.03%)	0.563
OMIM Disease Gene Loci (3,737)	1,658 (44.37%)	20,589 (4.93%)	0.000	1,664 (44.53%)	19,396 (4.96%)	0.000
OMIM Disease Gene Exons (3,737)	367 (9.82%)	458 (0.11%)	0.999	383 (10.25%)	492 (0.13%)	0.999
Autosomal Dominant Gene Loci (316)	247 (78.16%)	2,773 (0.66%)	0.023	245 (77.53%)	2,593 (0.66%)	0.031
Autosomal Dominant Gene Exons (316)	60 (18.99%)	70 (0.02%)	0.999	64 (20.25%)	78 (0.02%)	0.999
Autosomal Recessive Gene Loci (472)	386 (81.78%)	3,931 (0.94%)	0.065	402 (85.17%)	3,749 (0.96%)	0.009
Autosomal Recessive Gene Exons (472)	58 (12.29%)	78 (0.02%)	0.999	86 (18.22%)	109 (0.03%)	0.999
Cancer Disease Gene Loci (363)	301 (82.92%)	4,202 (1.01%)	0.651	307 (84.57%)	3,899 (1.00%)	0.821
Cancer Disease Gene Exons (363)	66 (18.18%)	85 (0.02%)	0.999	71 (19.56%)	98 (0.03%)	0.999
Dosage Sensitive Gene Loci (145)	120 (82.76%)	2,995 (0.72%)	0.604	125 (86.21%)	2,794 (0.71%)	0.728
Dosage Sensitive Gene Exons (145)	39 (26.90%)	51 (0.01%)	0.999	41 (28.28%)	58 (0.01%)	0.999
Genomic Disorders (52)	50 (96.15%)	14,178 (3.40%)	0.999	51 (98.08%)	13,373 (3.42%)	0.996
Pharmacogenetic Gene Loci (186)	97 (52.15%)	853 (0.20%)	0.517	96 (51.61%)	838 (0.21%)	0.105
Pharmacogenetic Gene Exons (186)	21 (11.29%)	27 (0.01%)	0.998	23 (12.37%)	29 (0.01%)	0.984
Imprinted Gene Loci (59)	39 (66.10%)	405 (0.10%)	0.989	37 (62.71%)	378 (0.10%)	0.982
Imprinted Gene Exons (59)	13 (22.03%)	15 (0.00%)	0.998	11 (18.64%)	13 (0.00%)	0.999
MicroRNAs (685)	8 (1.17%)	9 (0.00%)	0.785	11 (1.61%)	9 (0.00%)	0.836
GWAS Loci (419)	415 (99.05%)	9,413 (2.26%)	0.000	416 (99.28%)	8,852 (2.26%)	0.000
GWAS SNPs (419)	1 (0.24%)	1 (0.00%)	0.786	2 (0.48%)	2 (0.00%)	0.810
CpG Islands (14,867)	287 (1.93%)	1,516 (0.36%)	0.999	299 (2.01%)	1,508 (0.39%)	0.999
DNAseI Hypersensitivity Sites (95,709)	6,524 (6.82%)	7,165 (1.72%)	0.999	6,392 (6.68%)	6,914 (1.77%)	0.999
Recombination Hotspots (32,996)	16,839 (51.03%)	30,315 (7.27%)	0.000	16,211 (49.13%)	28,407 (7.27%)	0.000
Segmental Duplications (51,809)	17,172 (33.14%)	13,864 (3.32%)	0.999	16,518 (31.88%)	13,177 (3.37%)	0.999
Ultra-conserved Elements (481)	2 (0.42%)	2 (0.00%)	0.999	2 (0.42%)	2 (0.00%)	0.999
Affy 6.0 SNPs[h] (907,691)	1,556 (0.17%)	389 (0.09%)	0.999	3,022 (0.33%)	934 (0.24%)	0.999
Illumina 1M SNPs[i] (1,048,762)	2,318 (0.22%)	601 (0.14%)	0.999	4,789 (0.46%)	1,536 (0.39%)	0.999

*This table shows how structural variation affects different functional annotations and sequence characteristics in the HuRef genome. The leftmost column shows the names and total number of genomic features. The rest of the table is divided between gains and losses. Within the gain category, the first left column shows the number of (and percentage of total) genomic features impacted, and the second column shows the corresponding number of (and percentage of total) gain variants, and the last column shows the significance of the overlap as determined by simulations. An identical format is used for the losses.

[a] See Table S12 for a list of data sources.

[b] Based on a non-redundant list of 417,206 gains and insertions detected in this and the Levy *et al.* (Levy, et al., 2007) study of the HuRef genome.

[c] Based on a non-redundant list of 390,973 deletions detected in this and the Levy *et al.* (Levy, et al., 2007) study of the HuRef genome.

[d] Genes where a structural variant resides anywhere within the transcript (exonic and intronic).

[e] Genes from RefSeq data set where the entire transcript locus is encompassed by the structural variant.

[f] Genes from the RefSeq data set where exonic sequence is impacted by the structural variant. The non-redundant number of genes altered in some way by duplications and deletions is 4,867.

[g] Structural variants which overlap/impact a stop codon from the RefSeq gene set.

[h] Probes on the Affymetrix 6.0 Commercial array.

[i] Probes on the Illumina 1M array.

Currently, direct-to-consumer (DTC) testing companies and genome-wide association studies (GWAS) mainly use microarray-based SNP data (Fox, 2008; Ng, et al., 2009), but structural variants are typically not considered. HuRef indels/CNVs, however, overlap with 4,565 and 7,047 of SNPs on the Affymetrix SNP-Array 6.0 and Illumina-BeadChip 1M products (two commonly used arrays) potentially impacting genotype calling, most notably when deletions are involved.

Moreover, imputation of structural variation calls using tagging-SNPs captured 308/405 (76.0%) of the HuRef bi-allelic structural variants for which genotypes could be inferred (Appendix Table 18) (Conrad, et al., 2010b). Based on population data, rare structural variants with minor allele frequency$\leq$0.05 showed the lowest correlation with surrounding SNPs, thus indicating that these structural variants were least imputable (Figure II. 9). The fraction of imputable structural variants will be even lower when multi-allelic and complex structural variants are considered because the new mutation rate at these sites is higher.

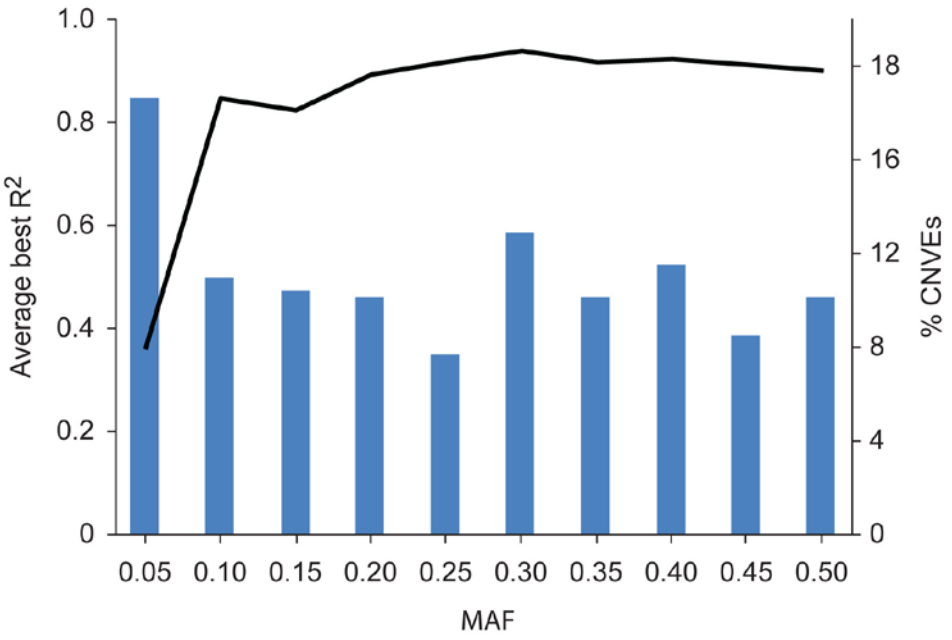


Figure II. 9. Tagging pattern for HuRef structural variants as a function of its minor allele frequency (MAF).
Linkage disequilibrium is depicted as the best r^2 between a structural variation and a HapMap SNP in 120 Europeans (CEU). There were a total of 405 bi-allelic polymorphic structural variation sites of overlap between GSV and HuRef loci; 24% of the structural variation loci have a HapMap SNP with $r^2<0.8$ in CEU, a cutoff below which HuRef CNVs would not be imputed simply by SNP detection. The line graph corresponds to the left y-axis, while the bar graph corresponds to the right y-axis. It should be noted that this analysis is performed on a small subset of bi-allelic structural variants and that the ability to impute a larger fraction of structural variants based on common SNPs would be even lower.

II.D Discussion

Human geneticists have long sought to know the extent of genetic variation and here, in the most comprehensive analysis to date, this study presents the latest estimates of greater than 1% within an individual genome. Using multiple computational and experimental approaches, it substantially expands on the structural variation map initially constructed by Levy and colleagues; more than 80% of the total 48,777,466 structurally variable bases have not been reported from the original sequencing of the HuRef genome.

My study here differs from previous studies in many ways. My mate-pair approach makes use of multiple different clone insert sizes, ranging from 2kb to 37kb, and this enables me to detect a wide size range of variants compared to previous paired-end mapping focused studies (Kidd, et al., 2008; Korbel, et al., 2007; Tuzun, et al., 2005). As expected, my results show that using several libraries with different insert size leads to increased variation discovery. Furthermore, the long sequence reads used here increase alignment accuracy, and enable the identification of intra-alignment gaps. Using microarrays, I am able to identify large size variants that can be challenging to identify by sequencing.

Furthermore, my results highlight that each variation-discovery strategy has limitations and that no single approach can capture the entire spectrum of genetic variation, thus emphasizing the importance of applying multiple strategies in structural variation detection. Figure II. 8 shows that the variation distribution of other personal genome sequencing studies, which relied almost exclusively on NGS technology, is substantially lower than the HuRef annotation across many size ranges.

There are still some regions such as heterochromatin (Figure II. 7) and highly identical segmental duplication regions where all of the current approaches have limited detection capabilities. To prevent false-discovery, I have used stringent alignment criteria, excluded alignments to multiple high-identity sequences, and will therefore likely miss variants within or flanking these sequences. Insufficient probe coverage and low intensity ratio fold-change also prevent microarrays from capturing CNV of highly-repetitive sequences (e.g. Alu elements). As such, I suspect there will be more variants to be discovered, but their

ascertainment will require specialized experimental (Alkan, et al., 2009; Kidd, et al., 2008) and algorithmic (Chen, et al., 2009; Lam, et al., 2010; Lee, et al., 2008) approaches. Further increase in read-depth can yield new variants. Indeed, the greatest relative number of structural variants discovered in HuRef is in the 10kb size range (Figure II. 2), corresponding to the interval with the highest clone coverage (Levy, et al., 2007) (Table II. 1).

The importance of structural variation to gene expression (direct and indirect) (Stranger, et al., 2007), protein structure (Ng, et al., 2008), and chromosome stability (Baptista, et al., 2008; Higgins, et al., 2008) is being increasingly recognized in normal development and disease (Buchanan and Scherer, 2008; Feuk, et al., 2006a). At the same time I show that structural variants are (i) grossly under-represented in published NGS sequencing projects, (ii) not always imputable by SNP-based association, (iii) ubiquitous along chromosomes impacting all known functional genomic features, and (iv) often large, complex, and under negative or purifying selection (Conrad, et al., 2010b; Pinto, et al., 2007). Coupling these observations with conjectures that prophylactic decisions will be best informed by higher-penetrance rare alleles (Bodmer and Bonilla, 2008) and that common SNPs explain only a proportion of heritability (Maher, 2008), these evidence argue persuasively that structural variants should gain more prominence in genomic medicine.

My results present the most thorough estimate to date of the total complement of genetic variation across the entire size spectrum in a human genome. My findings indicate that, to date, NGS-based personal genome studies, despite having generated a significant amount of valuable genomic information, have captured only a fraction of structural variants, with substantial gaps in discovery at specific points along the size range of variation. My data indicate that structural variation-discovery is largely dependent on the strategy used, and presently there is no single approach that can readily capture all types of variation and that a combination of strategies is required. (Structural variation detection by NGS is discussed further in Chapter V.) The data also show that structural variation impact many genes that have been linked to human disease phenotypes, and that interpretation of this data is complex (Lee and Scherer, 2010). Current genotyping services offered in the personal genomics field do not always include screening for structural variants, and I find that interpretation of current SNP based screening may be significantly impacted by the existence of structural

variants. I also show that many structural variants will not be amenable to capture using imputation strategies from high density SNP data, arguing for direct detection of structural variants as a complement to SNP analysis.

CHAPTER III: MECHANISMS OF FORMATION OF STRUCTURAL VARIATION IN A HUMAN GENOME

Data from this chapter have been included in the following publication:

Pang AW, Migita O, MacDonald JR, Feuk F, Scherer SW. 2013. Mechanisms of formation of structural variation in a fully sequenced human genome. Human Mutation (Early online publication).

I performed the breakpoint sequence analysis, mechanism-assignment for each variant.

III.A Introduction

In a study of individual genome sequences, to understand the impact and implications of genetic variation, we need precise details of its nature and content. Even with advanced approaches to discovery, only a limited portion of structural variation found in whole genome studies has breakpoints refined at the nucleotide level (Conrad, et al., 2010a; Kidd, et al., 2010a; Lam, et al., 2010; Mills, et al., 2011b; Perry, et al., 2008). Knowledge of the precise start and end of a structural variant is essential to determine its functional impact, to estimate its formation of mechanism, and to design targeted genotyping assay.

The main difficulty in resolving breakpoints is that they are not readily revealed by microarrays or whole genome sequencing using short read technologies. It is difficult to map short reads to tandem repeat loci whose length may be longer than the reads themselves, thus preventing detection of microsatellite polymorphisms. Also, short insert sizes limit the ability to properly anchor paired-end reads within complicated regions containing blocks of segmental duplication. So during breakpoint refinement at such regions, reads capturing the junction signature will likely be obscured by surrounding noisy alignments. In a recent population-scale genome sequencing study, slightly over half (53%) of the detected structural variants could be mapped to nucleotide resolution (Mills, et al., 2011b). Junctions of structural variation discovered through methods that lack precise nucleotide resolution (e.g. read-depth and read pair analyses) can, in principle, be refined through overlapping calls detected by methods with greater precision (e.g. assembly comparison and split-read). In practice, however, there are notable differences among methods in terms of size and locations ascertained (Mills, et al., 2011b; Pang, et al., 2010), thus limiting the amount of overlap to be achieved. Moreover, co-localization of multiple breakpoints can further limit the ability to map, detect and resolve complex variation by short sequencing reads.

Having precise structural variation breakpoints can enhance the ability to investigate the mechanisms responsible for variant formation. For this task, I needed a comprehensive set of structural variants with precise junction information. One such data set is the catalog of genetic variation discovered in the first sequenced personal human genome, the HuRef genome (Levy, et al., 2007; Pang, et al., 2010).

The HuRef genome was sequenced using Sanger capillary technology, and the long mate-pair clone-end sequences were subsequently assembled into high quality scaffolds. The availability of long reads and scaffolds enable accurate alignments, which in turn can yield precise variant breakpoint information even across more complex loci. This unique information facilitates significant progress in characterizing the origin (Xing, et al., 2009) and functional impact (Ng, et al., 2008) of variation discovered in HuRef. There are 739 small indels detected in the HuRef exome affecting 607 genes, and most of these coding variants are common and likely to be functionally neutral (Ng, et al., 2008). In addition, I found 189 genes to be completely encompassed by large gains or losses, and 4,867 genes whose exons are impacted (Pang, et al., 2010). Moreover, 573 genes with Online Mendelian Inheritance in Man Disease annotation are affected by variation.

Nevertheless, even with a comprehensive variation dataset containing breakpoint information, a thorough annotation of formation mechanisms of structural variation remained to be done. In this study, I examined the mechanism of formation for 408,532 gains, 383,804 losses and 166 inversions, and noted a differential proportion of mechanisms according to size. Ligation and replication slippage were more prevalent for small variants, whereas larger structural variants ($\geq$ 1 kb) were more commonly associated with retrotransposition and non-allelic homologous recombination (NAHR). To my knowledge, this is the first attempt to impute mutational mechanisms based on a near complete catalog of structural variation. This study represents an improvement over previous surveys (Table III. 1, Figure III. 1) (Conrad, et al., 2010a; Kidd, et al., 2010a; Lam, et al., 2010; Mills, et al., 2011b; Perry, et al., 2008), and I believe that this current work provides the closest estimate of the true proportions of various mutational processes.

Table III. 1. Summary of events and inferred mechanisms in current and two previous studies.

	Current Study Total sample size = 1			Kidd, et al., 2010* Total sample size = 17		Mills, et al., 2011** Total sample size = 185		
	Gains	**Losses**	**Inversions**	**Gains and losses**	**Inversions**	**Gains**	**Losses**	**Inversions**
VNTR†	109,297 (26.83%)	104,267 (27.26%)	2 (1.71%)	30 (3.08%)	0	122 (5.04%)	245 (3.76%)	0
REI	650 (0.16%)	892 (0.23%)	0	200 (20.55%)	0	1,994 (82.36%)	272 (4.18%)	0
NAHR	697 (0.17%)	1,289 (0.34%)	64 (54.70%)	219 (22.51%)	56 (69.14%)	226 (9.33%)	1,496 (22.97%)	0
NH	296,721 (72.84%)	276,062 (72.17%)	51 (43.59%)	524 (53.85%)	25 (30.86%)	79 (3.26%)	4,500 (69.09%)	0
Total	407,365	382,510	117	973	81	2,421	6,513	0

* as reported in Table 1 in Kidd, et al., 2010 study, which is based on the combined variation results of 17 samples. This project was comprised of Sanger-based sequencing of 13.8 million fosmid clones covering ~ 4.6 % of the human genome, thus explaining the comparatively low number of structural variants characterized.
** as reported in Figure 4 in Mills, et al., 2011 study, which is based on the combined variation results of 185 samples. This project was based mainly on low-pass short-read next generation sequencing data, thus explaining the comparatively low number of structural variants characterized.
† includes structural variants associated with microsatellites and minisatellite

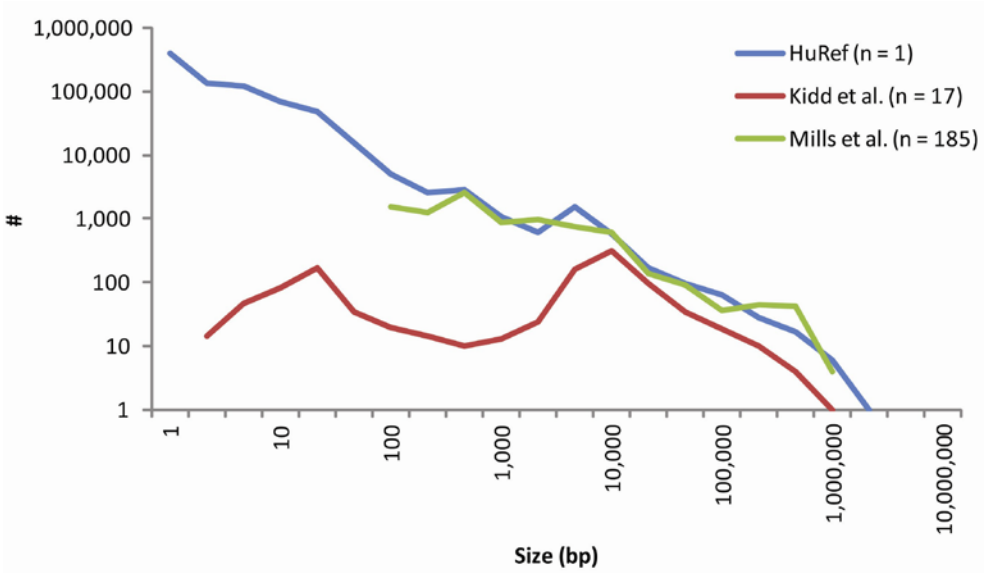


Figure III. 1. A comparison of the size and number of variants in three studies whose mutational mechanisms have been annotated.
The number of samples analyzed in each study is indicated next to the study name. Data from Kidd et al. study was obtained from Supplementary Table 2, while Mills et al. study from Supplementary Table 11 of the respective studies.

III.B Materials and methods

All non-SNP variation detected in HuRef from two previous studies (Levy, et al., 2007; Pang, et al., 2010), except those annotated as heterozygous mixed sequence variants (n = 21,480), were analyzed to impute their formation mechanism. In addition, 153 structural variants were sequenced from the larger size spectrum to add more loci. In total, I studied 408,532 gains, 383,804 losses and 166 inversions in HuRef, relative to the public reference assembly Build 36.

I classified the following four mechanisms in our pipeline: A) switch-over of intra- or inter-chromosomal homologous sequences by NAHR during repair or meiosis; B) ligation of double strand breaks (DSBs) by NHEJ or MMEJ; C) strand slippage or template switching at a replication fork that is stalled (FoSTeS) or broken (MMBIR); and D) transposition of retrotransposable elements. Furthermore, I distinguished variants located within highly mutable tandem repeat loci (i.e. VNTRs), and annotated them as copy number change of microsatellites and minisatellites. Such repeats have propensity to undergo copy number changes by recombination or replication slippage. Finally, because there is no obvious difference in the length of sequence microhomology among NHEJ, MMEJ, FoSTeS or MMBIR, I annotated them as non-homologous processes.

Specifically, for a variant to be classified as a micro- or minisatellite repeat, it must be annotated by the Tandem Repeat Finder program (Benson, 1999). The minimum unit size of a microsatellite repeat is 1 base pair (bp), and the minimum unit size of a minisatellite repeat is 10 bp. The repeat must begin outside the variant and extend into the variant covering at least 50% of its span. To identify NAHR variants, I first looked for extensive nucleotide homology (> 200 bp) flanking the variation breakpoints, based on the segmental duplication track from the University of California, Santa Cruz (UCSC) database and the RepeatMasker annotations, and then searched for homology of size $\geq$ 20 bp using the software Vmatch (www.vmatch.de) (with parameters -d -p -l 20 -identity 100). For REI, more than 70% of an indel had to be annotated by RepeatMasker (Smit, 1996-2010) as an L1, Alu or SVA element. For the remainder of variants with precise boundary information, I further searched for signatures of non-homologous mechanisms such as NHEJ and MMBIR (Hastings, et al.,

2009). I extracted 20 bp of flanking sequences from both the HuRef assembly and the NCBI reference assembly, build a local BLAST database based on the NCBI sequences, aligned the HuRef sequences to the database using BLAST (blastall -W 11 -g T -F F -S 3 -e 20) (Altschul, et al., 1990). By aligning the breakpoint flanking sequences in both assemblies, I aimed to identify DNA sequences present in the HuRef DNA and not in the NCBI assembly. Identifying regions of microhomology surrounding variants (< 20bp) was determined by running a custom PERL script Figure III. 2.

Using a 10 kb sliding window scheme, I screened for clusters of breakpoints of variation at least 1 kb in size. I compared the observed breakpoint density against a null model generated by simulations. While maintaining the size of variants, HuRef calls ≥ 1 kb were randomly shuffled along the chromosomes, and then I recorded the number of breakpoints observed in each 10 kb window. This simulation procedure was then repeated 1,000 times. To identify a candidate window that harbored a complex variation, I required that it must contain more real breakpoints than shuffled breakpoints in all 1,000 simulations. Finally, I further undertook manual inspection before annotating a region as having a complex variation event.

III.C Results

III.C.1 Mechanism of structural variation formation

Previous studies (Levy, et al., 2007; Pang, et al., 2010) identified 792,502 structural variants in the HuRef genome. I had breakpoint information for 406,963 gains, 382,196 losses and 88 inversions, and the majority of these calls were small (median size of 1 bp). An additional 153 larger variants were sequenced to obtain additional regions. In total, there are 789,340 structural variants (407,038 gains, 382,206 losses and 96 inversions) mapped at the base-pair level.

I applied my computational pipeline (Figure III. 2), and was able to assign the formation mechanism for 407,365 (99.71 %) gains, 382,510 (99.66 %) losses, and 117 (70.48 %) inverted sequences. For the remaining calls, I had insufficient breakpoint precision to confidently assign a mechanism. Non-homologous processes were associated with the majority of variants with a gain or loss of DNA (Figure III. 3 A and B), whereas NAHR was the dominant mode of genesis for inversions (Figure III. 3 C). Overall, 54.7% of inversions

were flanked by homologous sequences in opposite orientation, and the majority of those were mediated by large segmental duplication or L1 elements. The two largest NAHR inversions were L1- and duplication-associated (87,609 bp and 68,145 bp, respectively).

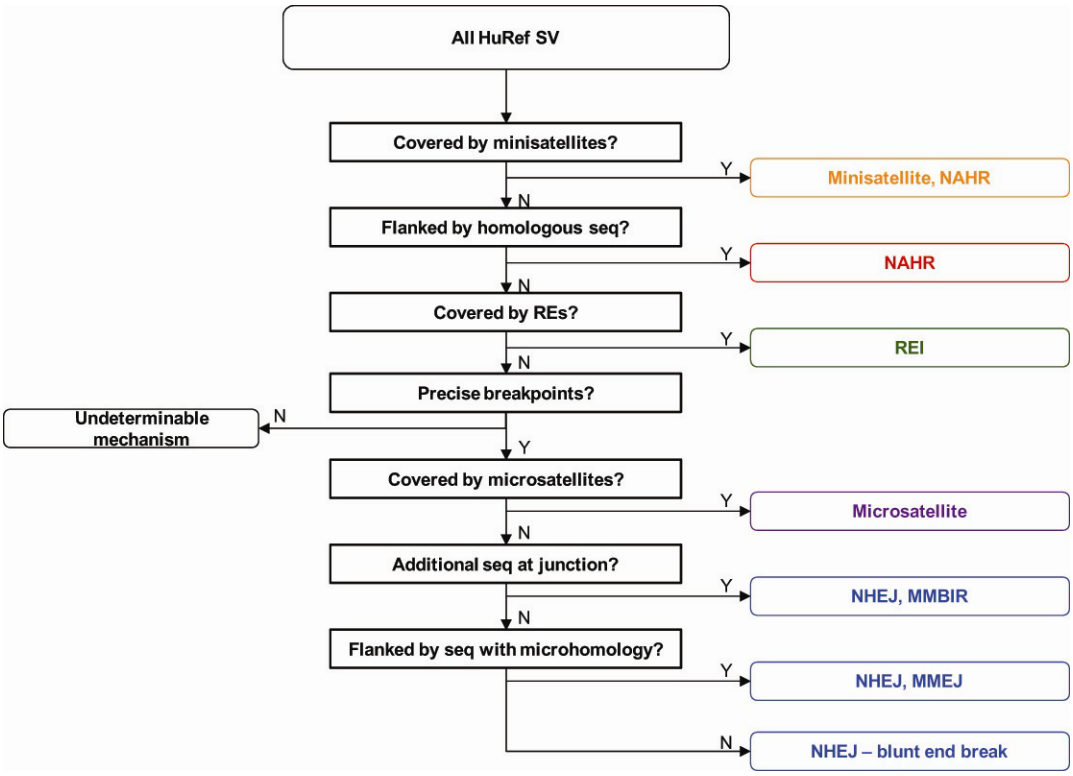


Figure III. 2. Mechanism assignment pipeline.
The computational pipeline in assigning variation formation mechanism is shown. Assignment of minisatellite, NAHR and REI does not require precise junction information, but such information is essential to assign the remaining categories. Note that the resulting mechanism assignment will change if the order of the analysis is rearranged. The primary reason of the order here is that I need to separate VNTRs from flanking microhomologous or homologous sequences, otherwise most of the expansions/contractions of tandem repeats would be incorrectly assigned as NHEJ or NAHR. Therefore, I decided to identify VNTRs before performing any flanking homology search. On the other hand, the number of REIs should be robust, and should not be affected by the ordering of the pipeline.

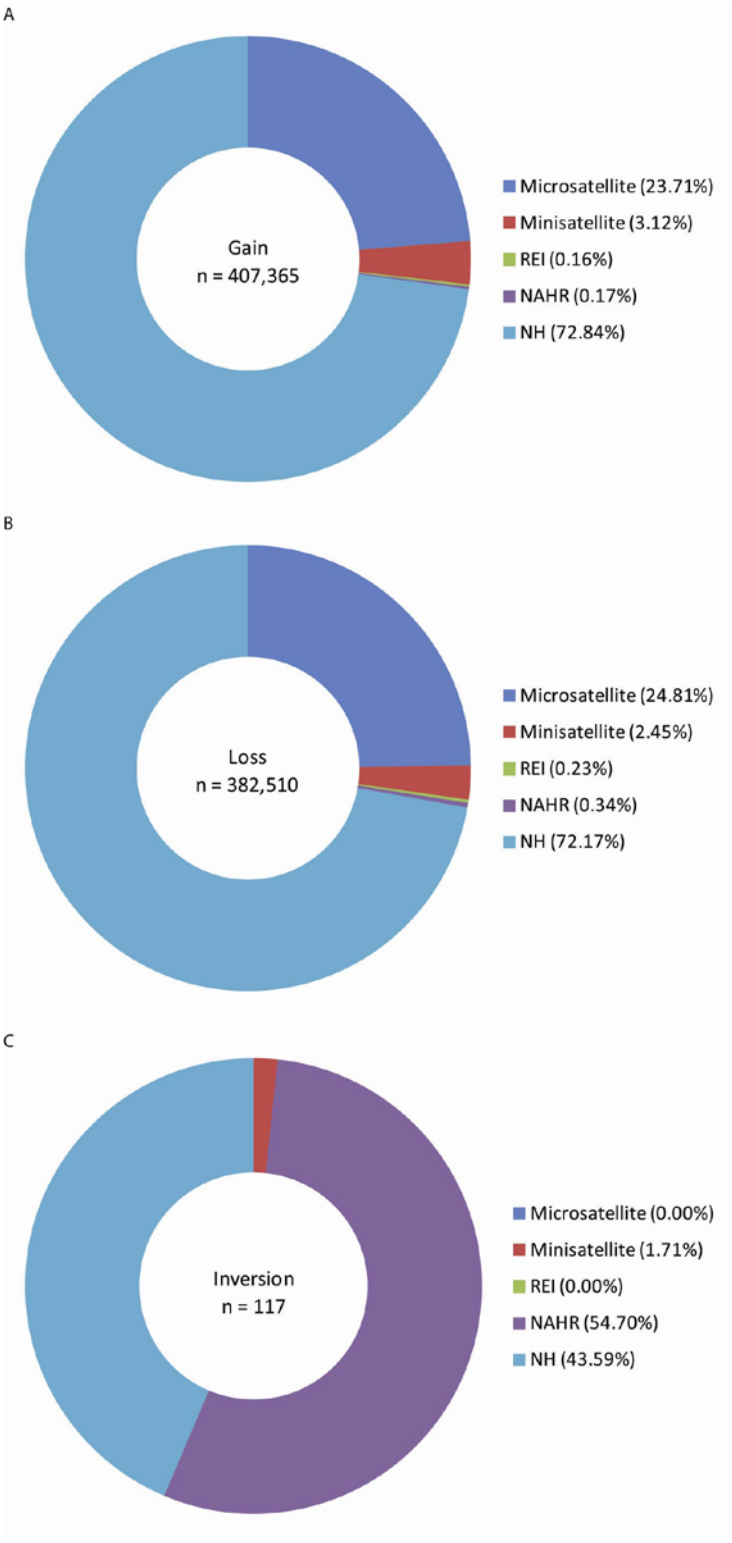


Figure III. 3. Relative proportion of mechanism of structural variation formation.
This panel of figures shows an overall proportion of mechanism irrespective of variation size. (A) Gain. (B) Loss. (C) Inversion. NH represents non-homologous processes.

I next investigated the relative proportion of formation processes across variant sizes. I found that NAHR was a dominant mechanism for all large gains, losses and inversions (> 10 kb). For gains and losses, I observed a gradient in the relative proportion of processes (Figure III. 4). The majority of small indels (< 10 bp) did not have any noticeable sequence signatures (Figure III. 4 A and B; Section III.C.1.iii). Most of the remaining gains and losses (up to 1 kb) were associated with micro- and minisatellites. Retrotransposition of Alu, L1 or SVA elements accounted for 20.96 % of the variation in the 100 bp to 1 kb range and 24.98 % of 1 kb to 10 kb variants. See Section III.C.1.i. Again, recombination errors, thus NAHR, was responsible for large structural changes (Section III.C.1.ii). Inversions showed a similar trend of changing mutational mechanism; non-homologous processes were responsible for most inversions of less than 1 kb, while NAHR was associated with the majority of larger variants (Figure III. 4 C).

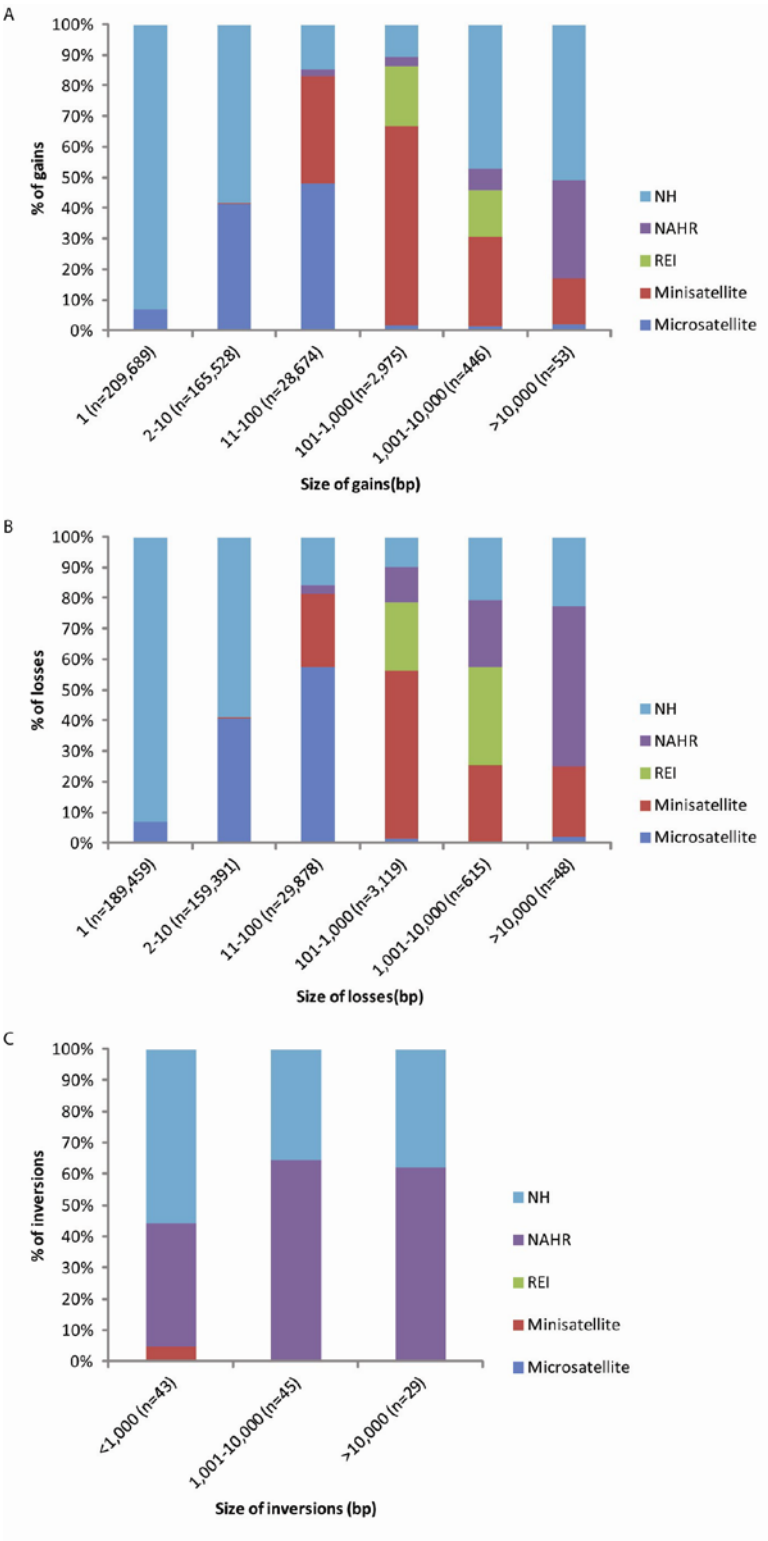


Figure III. 4. Relative proportion of mechanism divided by variant size.
The relative proportion of mechanism of variants of different length is shown. Generally, different mechanisms are responsible in forming variants of different size. (A) Gain. (B) Loss. (C) Inversion. NH represents non-homologous processes.

III.C.1.i Short tandem repeats and retrotransposable repeats

There were 213,564 (27.0 %) insertions, duplications and deletions that were associated with simple tandem repeats: homopolymer, micro-, minisatellite sequences. Changes in length and copy number are believed to be caused by slippage of simple repeats in recombination and replication (Richard and Paques, 2000). Interestingly, there were 2,653 insertions that reside in simple repeat loci, yet their sequences do not have the same base composition as the surrounding repeats. I did not consider these sequences as tandem repeats, and it would be incorrect to classify this category of mechanism solely by the genomic location while ignoring the nucleotide content. Instead, 90 % of these insertions are classified as formed by nonhomologous process. Upon examining the location of minisatellite variants, I noticed that they were clustered overwhelmingly near the end of chromosomes (Figure III. 5), which are some of the most dynamic region in our genome, displaying hypervariability with a large number of alleles.

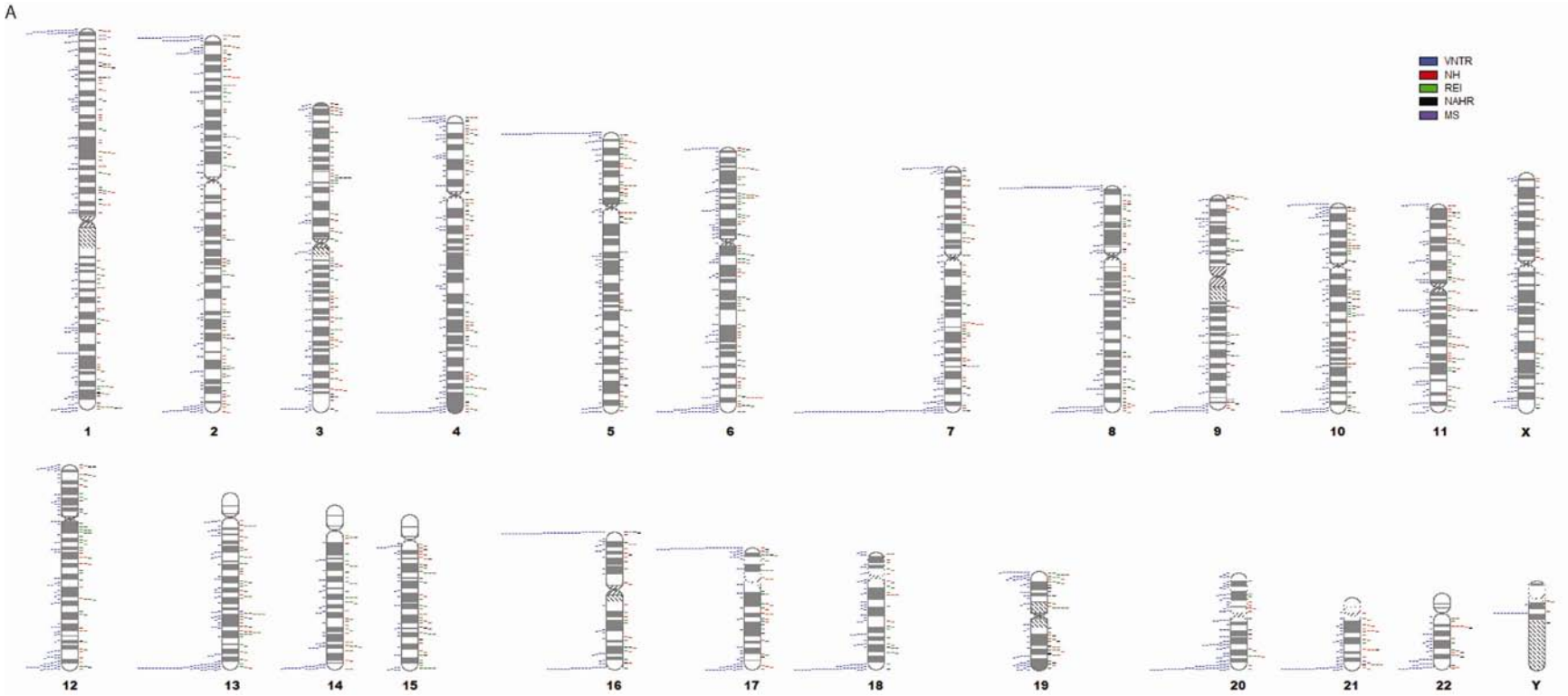


Figure III. 5. Ideograms illustrating the location of variants greater than 100 bp.
(A) Gain. (B) Loss. (C) Inversion. Note that VNTR represents the variation associated with minisatellite, and these variants are clustered at the end of chromosomes. NH denotes variants formed by non-homologous processes, REI by retrotransposable elements insertions, NAHR by non-alleleic homologous recombination, and MS as associated with microsatellites.

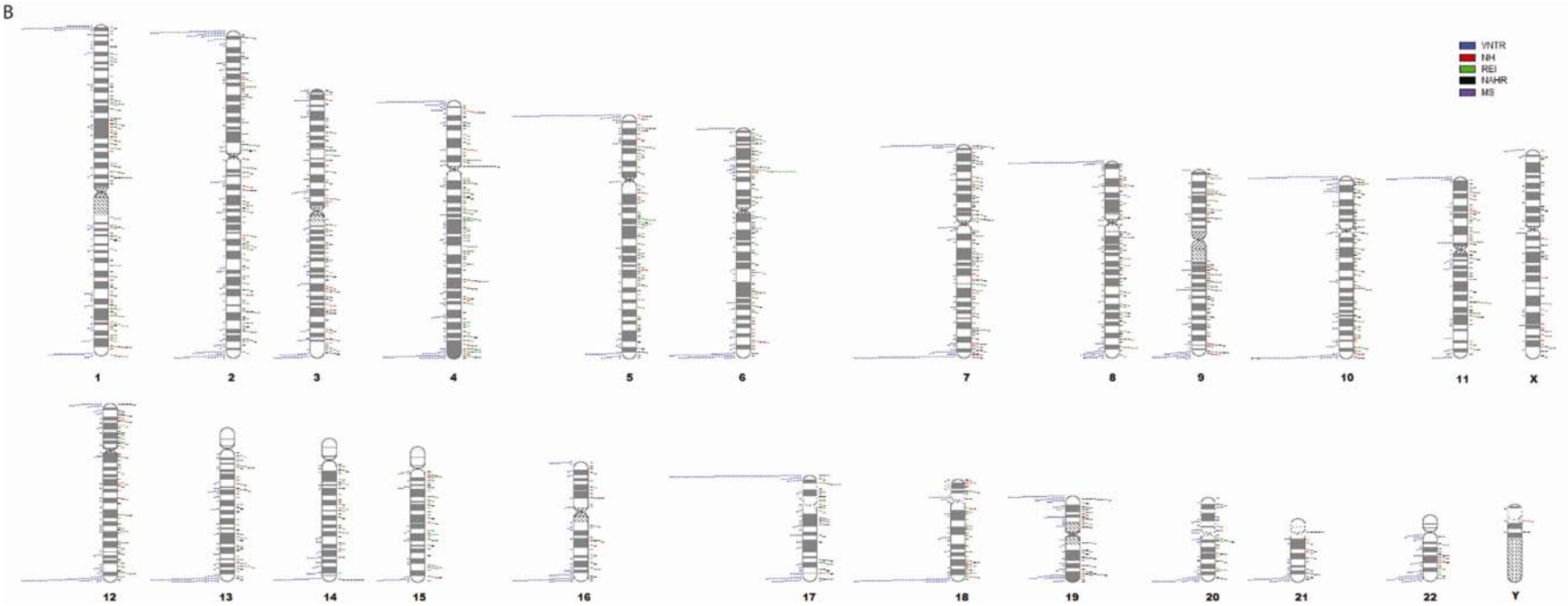


Figure III. 5. Ideograms illustrating the location of variants greater than 100 bp.
(A) Gain. (B) Loss. (C) Inversion. Note that VNTR represents the variation associated with minisatellite, and these variants are clustered at the end of chromosomes. NH denotes variants formed by non-homologous processes, REI by retrotransposable elements insertions, NAHR by non-alleleic homologous recombination, and MS as associated with microsatellites.

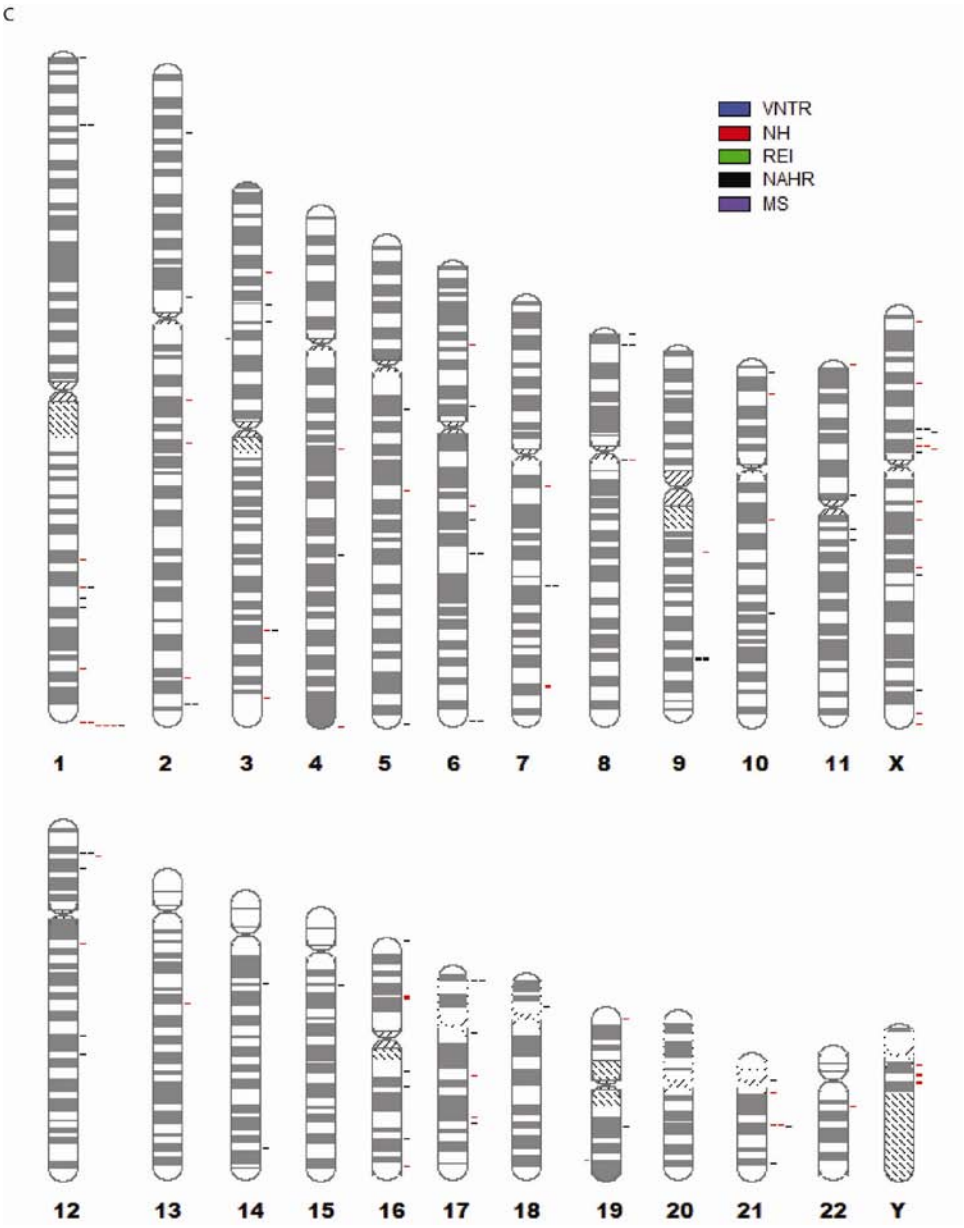


Figure III. 5. Ideograms illustrating the location of variants greater than 100 bp.
(A) Gain. (B) Loss. (C) Inversion. Note that VNTR represents the variation associated with minisatellite, and these variants are clustered at the end of chromosomes. NH denotes variants formed by non-homologous processes, REI by retrotransposable elements insertions, NAHR by non-alleleic homologous recombination, and MS as associated with microsatellites.

There were 1,542 (0.2 %) indels classified as non-long terminal repeat (non-LTR) retrotransposition. And consistent with previous reports (Lander, et al., 2001; Stewart, et al., 2011; Xing, et al., 2009), *Alu* elements are the most numerous transposable element in the human genome, constituting 1,045 events in the HuRef DNA. In addition, 96 gains and losses belonged to L1 retrotransposition and 17 to SVA, while 384 variants were associated with multiple Alu or L1 elements. L1 elements are the only currently known autonomous retrotransposons still active in humans (Konkel and Batzer, 2010). As expected, Figure III. 6 shows a "U" shape size distribution of L1 variants. There were 42 variants greater than 6 kb, the size of a full-length L1, and they may have maintained the ability to retro-transpose autonomously.

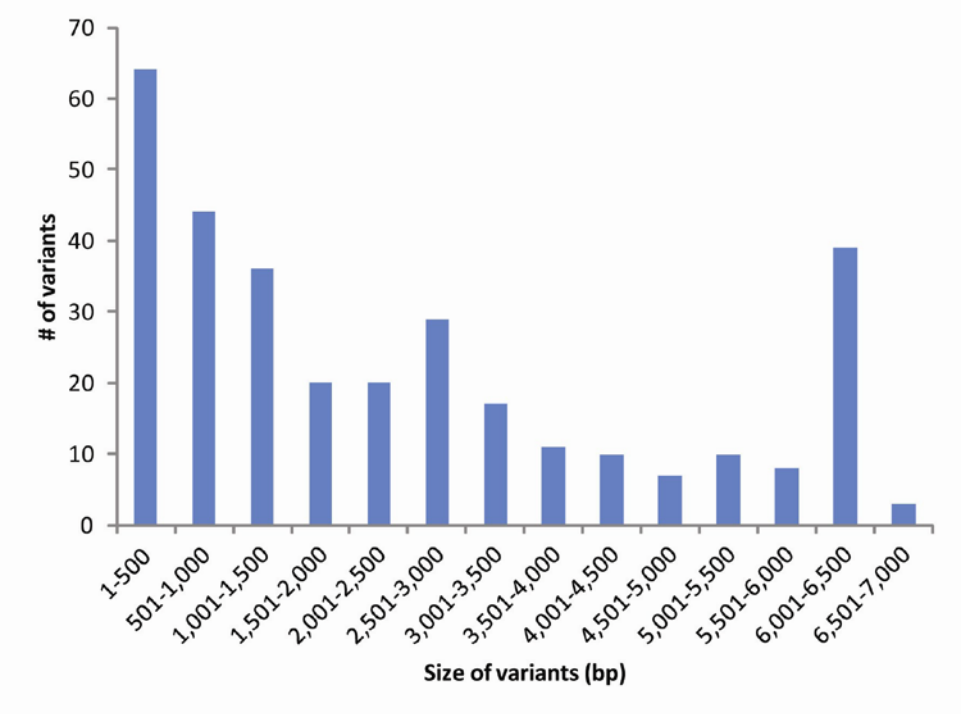


Figure III. 6. L1-associated variant size distribution.

III.C.1.ii Non-allelic homologous recombination

I examined stretches of homologous sequences (≥ 20 bp) surrounding variation junctions, as these sequences could have mediated meiotic chromosome/chromatid ectopic pairing. Specifically, I looked for large homologous sequences such as segmental duplications, medium size repeats such as LINE and SINE, and short perfectly identical nucleotides flanking each structural variation. I found that NAHR was responsible for 697 (0.2 %) gains, 1,289 (0.3 %) losses and 64 (54.7 %) inversions. In particular for inversions, the median distance was 1.9 kb between homologous copies, and the length of homology is 2.9 kb. In general, I noticed that there was a moderate but significant correlation between the size of variants and the length of their flanking homologous sequences (Spearman's correlation coefficient rho = 0.52. Besides length, homologs of high nucleotide similarity were better at mediating NAHR events. Of the variants surrounded by large segmental duplications, the majority (62.5 %) of them were accompanied by homologs sharing at least 95 % sequence identity.

III.C.1.iii Non-homologous processes

I next attempted to classify variants associated with replication and ligation. I searched for homologous sequences, and selected a threshold of 20 bp. This value was similar to the length of 34 bp of the "minimal efficient processing segments" required to mediate NAHR events between human alpha-globin genes (Lam and Jeffreys, 2006), and was the same as the threshold applied in a previous fosmid-sequencing study (Kidd, et al., 2010a). Of course, I could not rule out that there may be other currently unknown molecular mechanisms at work besides NAHR and non-homologous processes. In any case, I could only characterize variants whose nucleotide-level breakpoint has been resolved to be associated with non-homologous mechanisms. These represent variants detected by assembly comparison and split-read methods; and the variants with refined breakpoints.

I characterized the sequence content at each break by determining the number and content of nucleotides inserted in the break, and the amount of microhomology at the break. There were 9,898 (2.6 %) deletions and five (4.3 %) inversions with 1 to 10 bp of inserted sequence at breakpoints. With long reads and assembled contigs, I identified deletions as small as five bp

that had additional inserted sequence, and such sequence might correspond to non-template bases added as a consequence of imperfect NHEJ repairs. There were 117,505 (28.9 %) gains, 106,629 (27.9 %) losses and 28 (23.9 %) inversions that showed 1 – 20 bp of flanking homology, and this signature would indicate the formation processes to be either NHEJ, MMEJ, FoSTeS or MMBIR. However, it should be noted that the annotation of one to three base pairs of microhomology could be the result of random chance, or could be false positives due to sequencing-, alignment- or assembly error. After their exclusion, the remaining 4,370 insertions and 5,008 deletions were flanked by at least 4 bp of homologous sequences. Of those deletions displaying stretches of microhomology, 3,773 (3.5 %) also had insertion sequences at the breakpoint. Surprisingly, 14 (1.2 %) NAHR-associated deletions also had breakpoint-insertion sequences. The relative proportions were significantly different ($P < 1.00x10^{-4}$, chi-square test), thus substantiating the difference among homologous and non-homologous mechanisms. Finally, 179,216 (44.7 %) gains, 159,535 (41.7 %) losses and 18 (15.4 %) inversions were simple blunt-end junctions that had neither additional sequence nor microhomology. Figure III. 7 shows the distributions of signature size for microhomology, blunt-end and inserted sequence at the breaks of deletion variants. In general, the data shows that non-homologous mutational processes facilitated the formation of the vast majority of insertions (72.8 %) and deletions (72.2 %), but most of them (99.8 %) were no bigger than 100 bp.

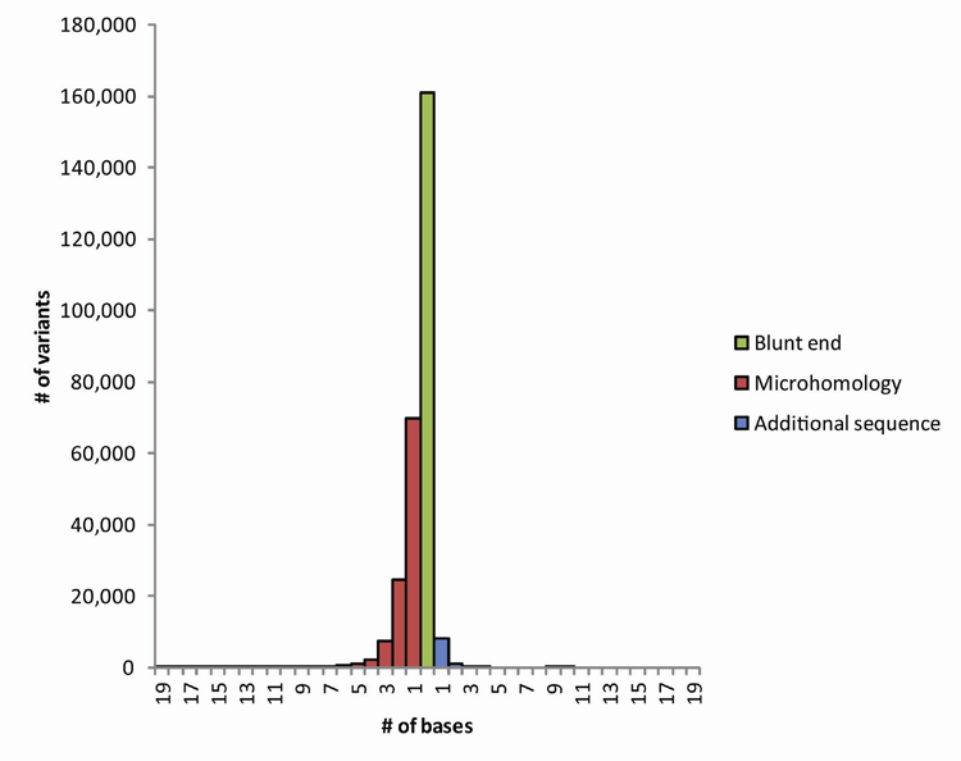


Figure III. 7. Distribution of deletions breakpoints with blunt end, microhomology, and additional sequence signatures.

III.C.1.iv Comparison with other mutational mechanism studies

This study is the first to examine the relative proportion of mechanism across the entire size spectrum of structural variation. Table III. 1 illustrates a comparison of the results of the current study and two published studies (Kidd, et al., 2010a; Mills, et al., 2011b), which examined aggregated data across multiple genomes. While thorough in the analysis, these studies only annotated indels or variants of a certain size (Table III. 1, Figure III. 1). With a more complete variant set, the numbers of microsatellite and variants formed by non-homologous processes were greater than in previous estimates. Furthermore, the proportion of NAHR was lower than previously found, as this process was relevant for the few but large variants. Overall, differences in the number and proportion of each mechanism category between this data and others reflect the underlying differences in variant ascertainment and annotation, and the number of samples typed. I benefit from the availability of a variation set of all types and of a wide range of sizes, so I can better approximate the true proportion of mechanism operating in the genome.

III.C.2 Complex variants

Hastings and colleagues (Hastings, et al., 2009) propose that template switching during replication is responsible for the formation of complex variants. Complex variants are those that have more than one simple rearrangement, and have two or more breakpoint junctions (Quinlan and Hall, 2012). Of course, it is also possible that such architecture came about via multiple independent simple events. Currently, to my knowledge, there is no definitive sequence signature that can be used to identify complex events. Nonetheless, I attempted two approaches to screen for clusters of local variants that potentially arose by a single mutation event.

I first examined whether there were insertion sequences within larger deletions or inversions. All of the breakpoint insertion sequences were short (< 10 bp): too small to discern whether they originated from distinct genomic loci brought about by template switch during DNA replication. (See Section III.C.1.iii). I next searched for multiple rearrangements at a single locus. I explored the HuRef genome to identify loci where multiple structural variation breakpoints were present. From simulations (see Section III.B), I established a null model of

breakpoint density, against which I compared my observation to ensure that the observed clusters were unlikely to have occurred by chance. After comparing to simulated data and with subsequent manual inspection, I identified 56 regions where distinct breakpoints of variants > 1 kb co-localized within 10kb.

Of the annotated loci, the most common pattern was consecutive deletion breakpoints, constituting 50% of the cases. The next most common pattern was adjacent insertions or duplications, accounting for 16 loci (28.57 %). I also observed other combinations: deletions and insertions embedded within an inversion; a triplication within a duplication next to two deletions; and a deletion embedded within an inversion contained in a duplicated region. I further genotyped the last region, and the results are described in Chapter IV.

I emphasize that some of these breakpoints represented genuine single complex variants, while others were the results of serial independent events. There were 23 out of 56 (41.07 %) multi-breakpoint clusters that may be an accumulation of independent events, as they overlapped with known segmental duplications. Aside from those, 12 complex events impacted exons, and 18 overlapped regions that lacked synteny with primate sequences (Table III. 2). Hence, these events may have evolutionary significance.

Table III. 2. List of 10 kb regions that show clustering of breakpoints of variants whose size is at least 1 kb.

Cluster Coordinates	# of breakpoints	Recombination hotspot	Segmental duplication	GC content	Genes	Synteny with primates
chr1:1210001..1220000	4	n	n	68.2386	SCNN1D+ACAP3	gap
chr1:245040001..245050000	3	n	y	46.5012	-	break
chr2:194390001..194400000	3	n	n	37.6402	-	-
chr2:202440001..202450000	3	y	n	41.902	CDK15	-
chr2:219760001..219770000	4	y	n	44.4614	-	break
chr2:242350001..242360000	3	n	y	60.9193	D2HGDH	gap
chr3:37720001..37730000	5	n	n	45.5319	ITGA9	-
chr3:48500001..48510000	4	y	n	44.9823	SHISA5	break
chr3:164030001..164040000	3	n	n	33.1409	-	-
chr3:196990001..197000000	4	n	y	57.8509	MUC4	break
chr4:48820001..48830000	4	n	y	40.7881	-	gap
chr4:189600001..189610000	5	n	y	43.1984	-	break
chr4:190840001..190850000	4	n	y	41.8098	-	break
chr4:191000001..191010000	4	y	n	44.3177	-	break
chr5:600001..610000	3	y	y	51.6031	-	gap in human
chr5:1120001..1130000	3	y	y	62.8091	SLC12A7	gap
chr5:49470001..49480000	3	n	n	39.4136	-	gap
chr5:51430001..51440000	3	n	n	37.4957	-	gap
chr5:90530001..90540000	5	y	n	36.0535	-	-
chr5:177750001..177760000	6	y	n	53.0854	COL23A1	break
chr6:310001..320000	5	y	n	49.2756	-	gap in human
chr6:31380001..31390000	5	n	n	44.998	-	-
chr6:31400001..31410000	3	n	n	40.4016	-	-
chr6:32600001..32610000	5	n	y	42.9543	HLA-DRB5	-
chr6:57400001..57410000	3	n	n	37.389	PRIM2	-
chr6:161120001..161130000	3	y	n	39.7957	-	-
chr6:168130001..168140000	3	y	n	51.3425	-	break
chr6:170320001..170330000	4	n	n	54.0094	-	break
chr6:170540001..170550000	3	n	n	46.8858	FAM120B	break
chr7:1880001..1890000	4	n	y	49.8606	MAD1L1	-
chr7:100430001..100440000	3	n	y	51.5118	MUC12	break

chr7:154080001..154090000	3	n	y	41.9201	DPP6	break
chr7:157630001..157640000	3	y	y	53.2654	PTPRN2	break
chr8:1320001..1330000	3	y	y	53.6512	-	break
chr8:1820001..1830000	3	n	n	46.6394	ARHGEF10	break
chr8:18490001..18500000	4	y	n	35.8669	PSD3	gap
chr8:39360001..39370000	3	y	n	28.6307	ADAM5P	gap
chr8:58280001..58290000	3	n	n	55.9449	-	break
chr9:135860001..135870000	4	n	y	56.2315	-	gap
chr10:27640001..27650000	3	y	y	38.0937	-	complex
chr10:135140001..135150000	4	y	y	48.0839	-	complex
chr11:1890001..1900000	3	n	n	57.4181	TNNT3	-
chr12:129690001..129700000	4	y	n	50.6697	-	-
chr12:131320001..131330000	5	n	y	59.865	GALNT9	gap
chr13:45930001..45940000	4	y	n	47.9043	-	break
chr13:98050001..98060000	4	y	n	37.3917	-	-
chr13:113860001..113870000	3	n	n	57.5622	-	-
chr14:105300001..105310000	3	n	y	62.5484	-	break
chr15:89780001..89790000	3	y	n	37.2457	-	gap
chr16:830001..840000	2	y	n	58.6445	-	-
chr16:83990001..84000000	3	y	y	55.328	-	gap
chr17:5530001..5540000	3	y	n	47.5331	-	gap
chr18:74890001..74900000	4	n	y	48.0441	-	break
chr19:58010001..58020000	3	y	y	46.4685	ZNF28	-
chr21:10110001..10120000	4	n	y	42.3705	BAGE3	-
chrX:78800001..78810000	3	n	n	32.7793	-	-

III.D Discussion

In this study, with the availability of breakpoint sequences, I have undertaken a comprehensive analysis to fully investigate the underlying mechanisms that contributed to the formation of all non-SNP variants discovered in a single individual. Overall, variants derived from non-homologous events – those associated with NHEJ, MMEJ, FoSTeS, and MMBIR – are the most prominent, constituting up to 72.84 % of gains and 72.17 % of losses. However, once I subdivided by variant size, I discovered that most of these were no larger than 10 bp. Similarly, most micro- and minisatellite associated indels were small, and formed by strand slippage during DNA replication. Furthermore, we noticed that REI and NAHR were more prominent with variants larger than 1 kb. Multiple mechanisms for variant formation had been recognized as operating in the genome, but this study has improved upon earlier estimates of their relative proportion (Table III. 1).

Figure III. 4, showing the relative proportion of mechanism by the variant size, clearly indicates a few notable features. The first is the division between small and large indels. They are often detected by different approaches, and have generally been treated as separate entities. Based on mechanism profiles, my results in Figure III. 4 support this distinction and would suggest a size dividing line of ~100 bp. Another notable feature is the abundance of evidence for events associated with non-homologous processes that appeared to be "random", that lacked any notable sequence signature or any correlation with known genomic features (Figure III. 5). Surely, some small non-homologous events – those not associated with short tandem repeats – are similar to SNPs, and are indeed distributed throughout the genome. Yet for the large variants, there may be other systematic explanations for their apparent random location. One such explanation may be chromatin spatial proximity and closeness in replication timing, as seen in cancerous alterations (De and Michor, 2011; Fudenberg, et al., 2011). Interestingly, a recent study correlates replication timing and structural variation mechanism, and shows that hotspots of NAHR-mediated variants are enriched in early replication regions of the genome, while variant hotspots associated with non-homologous processes are more enriched in late replicating regions

(Koren, et al., 2012). Examining sequence *in cis* alone is perhaps not sufficient to resolve mechanisms; additional *in trans* experiments may be needed to yield clarity.

I also attempted to identify complex rearrangements that consist of more than one simple variant. These complex structural variants can rearrange exons, shuffle regulatory elements or disrupt multiple genes and pathways (Carvalho, et al., 2009; Lee, et al., 2007; Zhang, et al., 2009b). While excellent at detecting simple structural variants, current bioinformatics tools do not recognize these difficult, yet important variants. Here, I created custom approaches to search for co-localization of breakpoints and for non-template sequences at junctions, and then manually inspected each candidate. These approaches are not feasible for population-based whole-genome sequencing studies, so automated programs for such purpose are needed.

Variation junction information is crucial for this project; however, there are still some calls (1,167 gains, 1,294 losses and 49 inversions) that cannot be properly annotated, as there is sufficient precision to identify nucleotide-level signatures such as flanking microhomology or non-reference additional sequences. These variants have been discovered by lower resolution microarrays or mate-pair mapping. Approaches that can discover variants at full resolution may also generate spurious results due to errors in assemblies or issues with alignments, further limiting accurate breakpoint assignments. In this data, there are five multi-breakpoint complex regions that contain inversions, and precise inversion junction data is available for four of the five (80 %). It is possible that complex loci may not be resolved solely by genome-wide approaches due to underlying sequence structures and technical limitations. Perhaps, their resolution will require traditional targeted approaches, or creative combinations of high-throughput methods such as sequence capture using probes/baits (Conrad, et al., 2010a) designed from variant junction libraries (Lam, et al., 2010).

In conclusion, with precise breakpoint information, I assigned formation mechanisms to structural variants from the entire size spectrum in the HuRef individual genome. I

demonstrated that different mechanisms are more prominent within different size classes. My study offers additional insights into the origin and complexity of genome variation.

CHAPTER IV: COMPLEX BREAKPOINT STRUCTURES ASSOCIATED WITH MICROSCOPIC INVERSIONS

Data from this chapter have been included in the following publication:

Pang AW, Migita O, MacDonald JR, Feuk F, Scherer SW. 2013. Mechanisms of formation of structural variation in a fully sequenced human genome. Human Mutation (Early online publication).

I performed some of the breakpoint PCR experiments together with Dr. Ohsuke Migita. Dr. Ohsuke Migita also performed qPCR experiments and analysis of the QIAxcel results. I performed the haplotype analysis.

IV.A Introduction

Inversions have traditionally been considered to be a form of balanced rearrangement presumably with no gain or loss of DNA (Kidd, et al., 2010a). Significant knowledge of inversions and translocations comes from cytogenetics experiments. Microscopic structural abnormalities rearrangements occur in about 1/375 live births, with about three quarters being balanced rearrangements (Nussbaum, et al., 2007). The risk of a serious congenital anomaly is estimated to be 9.4 % for inversions (Warburton, 1991).

The discovery of submicroscopic inversion, however, is rather modest compared to copy number changes, mostly due to the limited number of genome-wide tools. Many are identified in clinical cases, where inversions cause no apparent deleterious phenotype in parents but predispose subsequent rearrangements in offspring. For example, one third of the parents of patients with Williams-Beuren Syndrome have a 1.5 Mb inversion at 7q11.23 (Osborne, et al., 2001). Similarly, inversions at the olfactory receptor gene clusters on 4p16 and 8q23 are believed to mediate the recurrent t(4;8)(p16;p23) translocation by unusual meiotic exchanges, as the mothers of subjects with the de novo translocation all have heterozygous inversions on both 4p and 8q regions (Giglio, et al., 2002). Recent studies use assembly comparison across species (Feuk, et al., 2005; Khaja, et al., 2006) and mate-pair mapping (Kidd, et al., 2008; Tuzun, et al., 2005), and both approaches offer greater resolution in inversion discovery. Now it is known that inversions can suppress recombination between heterozygotes during meiosis, and can confer reproductive advantage (Stefansson, et al., 2005), and can drive evolutionary divergence (Feuk, et al., 2005).

Nonetheless, the number of polymorphic inversions identified is much less than indels and CNVs. As of September 2012, the DGV hosted 833,981 gain and loss entries contrasting to 906 inversions (Iafrate, et al., 2004; Zhang, et al., 2006). Inversions cannot be detected by genomic microarrays, and they are difficult to be found by mapping short reads generated by NGS. For instance, there is no inversion reported in a recent population sequencing study (Mills, et al., 2011b).

The HuRef inversion dataset is comparatively more complete than other personal genome datasets (Table I. 2). The HuRef assembly has been constructed from high quality and long Sanger-based sequences, thus yielding precise inversion breakpoints. A total of 166 inversions have been detected by two complementary approaches: assembly comparison and mate-pair mapping (Levy, et al., 2007; Pang, et al., 2010). To obtain a better understanding of the impact and origins of inversions in the human genome, I selected 8 HuRef junction-resolved inversions and genotyped these in human populations. I discovered that the structures of inversion could be complex, often accompanied by gains and losses of DNA, create conjoined genes, and their frequencies could exhibit population differentiation. Finally, I found inverted regions where the reference assembly may have been misassembled, or represents the minor human alleles.

IV.B Materials and methods

IV.B.1 Genotype analysis

PCR assays were designed to genotype eight HuRef inversions for which I had nucleotide breakpoint information; four had been detected by an assembly comparison method, and the other four were detected by mate-pair mapping and subsequently refined by breakpoint sequencing. These were chosen to represent different formation mechanisms from the previous chapter. Specifically, I selected five inverted loci that were formed by non-homologous processes, and three associated with NAHR. For the latter three, I designed PCR primers to amplify across the flanking homologous segmental duplications.

I selected four oligonucleotide primers for each region, with two primers outside the variant, and two within the inversion region. One of the two within the variant was based on the NCBI reference orientation, whereas the other one was based on the HuRef DNA orientation (Feuk, et al., 2005) (Figure IV. 1). All primers were optimized using a gradient hybridization temperature from 52 to 70 °C. The experiments were carried out using the Agilent Technologies (Santa Clara, California) PicoMaxx High Fidelity PCR System kit. The PCR cycling conditions were 95 °C for 5 min, followed by 30 cycles of (95 °C for 40 s, optimized annealing temperature for 40 s, 72 °C for 60 s per kb of product length), and a final extension 72 °C for 7 min. To estimate the frequency of the variant allele, I genotyped a panel of 42

human samples (10 HapMap Yoruba Nigerians (YRI), 10 HapMap Europeans (CEU), 10 HapMap Japanese (JPT), 10 HapMap Han Chinese, (CHB) NA15510, and HuRef), three chimpanzee, and one orangutan samples (Table IV. 1).

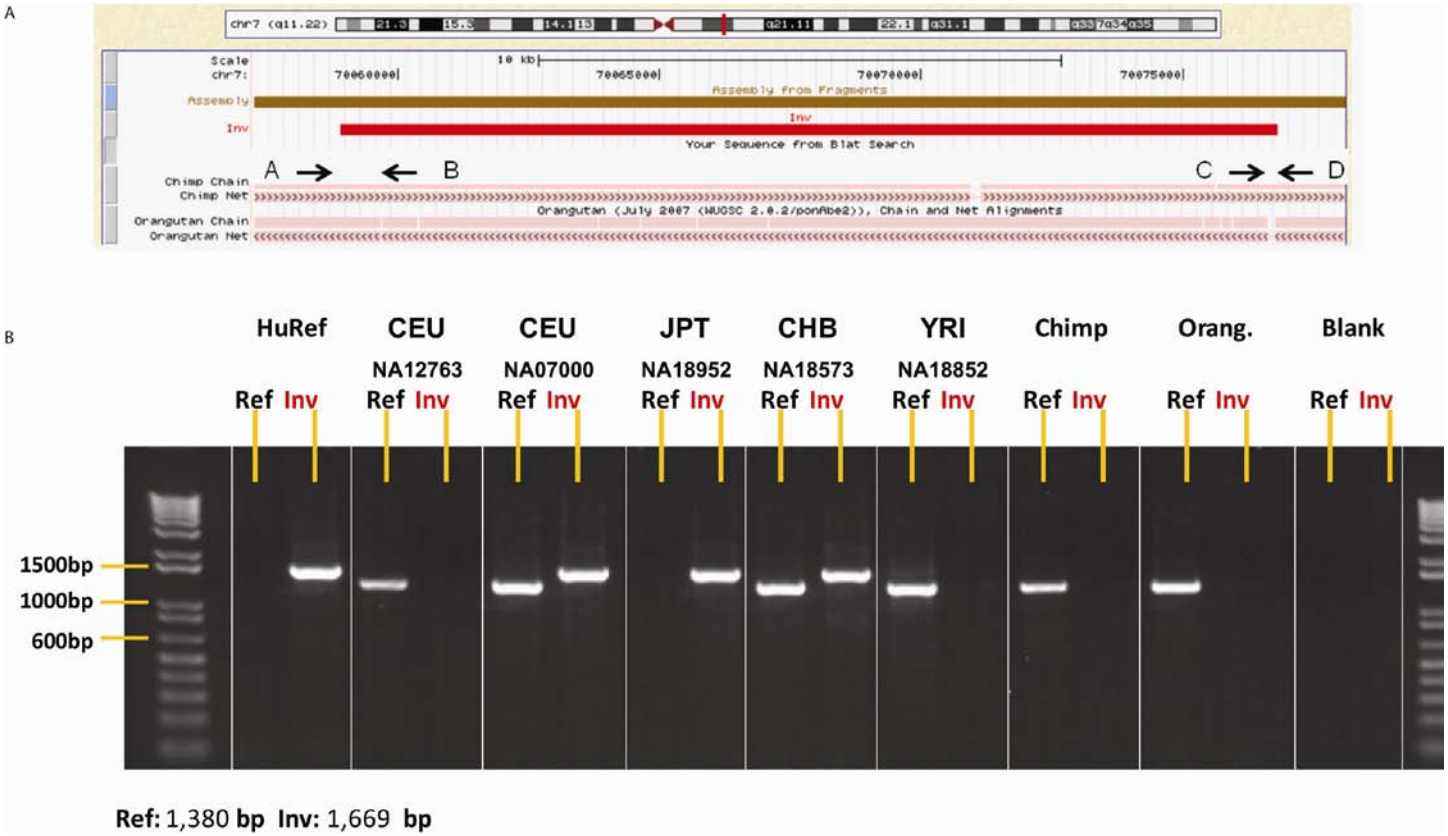


Figure IV. 1. PCR assay.
(A) Schematic diagram shows an example of a 17.9 kb inverted region at 7q11.22 as shown as the red track in the genome browser. Four primers A, B, C and D target the breakpoints. In the absence of inversion AB and CD sets will be amplified, whereas in the presence of inversion, AC and BD pairs will amplify. (B) A typical PCR result. This gel picture shows the PCR results for 9 samples with lanes loaded alternatively between the reference and inversion assays. The genotypes from left to right are as follow: HuRef homozygous for inversion; NA12763 homozygous for reference; NA07000 heterozgyous; NA18952 homozygous for inversion; NA18852 homozgyous for reference; the chimpanzee sample homozygous for reference; and the orangutan sample homozygous for reference.

Table IV. 1. List of variants and their primers used for inversion genotyping.

PCR primers							
Type	**Locus**	**Chrom**	**Start**	**End**	**Size**	**Primer label**	**Sequence**
inv	Xp11.3	chrX	46,695,748	46,715,632	19,885	A	TTCTGCCTGTGTAAAGGATGC
						B	GGAGCCAAAGGACTTGGTTT
						C	TGTCCACCTAACTGCACCAA
						D	ACCTCACTCGGTGGTCAACT
inv	7q11.22	chr7	70,058,905	70,076,823	17,919	A	AACAGGTTGAGGAAAGACCATC
						B	CTTCCTTCACAGACAGAACACG
						C	ATTGAATTAGTTGCCCATTTGC
						D	ATTCATTCCCTACACTGCATCC
inv	16q23.1	chr16	73,797,599	73,814,159	16,561	A (A3*)	TGACCTGGTGGAGTCTAGGG
						Ar	TCAGCATTCTGACCGTGAAC
						B (B3*)	TCGAGCCTCACCCTCTTAAA
						C	TCACTTCCTGCATGTTGACG
						C (C3*)	TGCCATTTTATGGTGTGGAA
						D	CAGTAAAGCTGGTTTGACCAATAG
inv	16q24.1	chr16	83,746,237	83,747,302	1,066	A	CACCTGGATGCCCACTTATT
						B	GATGGAGGTGCATTCGATTT
						C	AAATCGAATGCACCTCCATC
						D	TGGGTATATGGATGGGAGGA
inv	4q22.1	chr4	89,066,188	89,077,724	11,537	A	N/A
						B	GGAAACATGGGGATAAGAAACA
						C	TTAGGATTTGAACAAGGCCAGT
						D	GAGAGCTTCTGGCAGGCTTAC
inv	1q31.3	chr1	196,023,411	196,024,609	1,199	A	CTCAGGGACTTGGATTAACCTG
						B	GGCCCTTTTATCCTCCAATTAC
						C	TGCAAACTTTCTGGCTACTCTG
						D	N/A
inv	6q27	chr6	168,835,529	168,836,601	1,073	A	AACGTGGACGCGATACTACC
						B	ctggggaacaggacacaact
						C	agccagaagaagggaagagg
						D	CCATGCAGCTGCTTTTTACA

inv	3q26.1	chr3	164,008,436	164,030,337	21,902	A	TTGAAACCTCAGAGTTCCCATT
						B	TGTGCCAGTATTTGATCTCCAC
						C	AAAGAGACCCATTCTGCTTGAG
						D	N/A

Quantitative PCR primers

Type	Locus	Chrom	Start	End	Size	Forward primer	Reverse primer
del	3q23.1	chr3	164,008,296	164,030,349	18,936	ATGCCCTCATCAACAATGCTA	TTGTCTTTGGAGGCTGCTATTT
dup	3q23.1	chr3	163,994,833	164,109,038	114,206	ATTCCCAGGTCTTAGCCTTCTC	TAAGCCTTTCATCTTCCTTCCA

The inversions in all eight loci were genotyped using the above protocol and additional experiments were performed to better elucidate the structure of two regions. At 3q26.1, I reported a duplication, an inversion and a deletion overlapping one another in Chapter III. Here, in addition to genotyping the inversion, I also designed Life Technologies' SYBR Green based qPCR assays to test the duplication and deletion on the panel of samples (Carlsbad, California). Each assay was run in triplicate and the *FOXP2* gene was used as the internal control for relative quantifications (Feuk, et al., 2006b). The thermal profile for the qPCR was 95 °C for 5 min, followed by 40 cycles of (95 °C for 5 s, 60 °C for 11 s), followed by 95 °C for 60 s, 55 °C for 30 s, and finally 95 °C for 30 s. The primers used are listed in Table IV. 1.

Furthermore, to identify the population frequency of the 16q23.1 inversion/deletion impacting *CTRB1* and *CTRB2* genes, the DNA of 871 individuals from 57 populations from the HGDP-CEPH Human Genome Diversity Panel were genotyped (Cann, et al., 2002). To enable the genotyping this large panel of samples, I selected the QIAxcel instrument (Qiagen, USA), basing on capillary electrophoresis and employing a gel cartridge, to detect and size-measure PCR products. The QX Alignment Marker, which consisted of 15 bp and 5 kb bands, was injected into the cartridge with each 5 uL of PCR product, and this marker enabled the QIAxcel ScreenGel software to align the lanes automatically. I used the manufacturer recommended AM420 method in analyzing the PCR results.

IV.B.2 Haplotype analysis

Inversions were genotyped across ten HapMap samples per ethnicity, providing a sample size that was sufficient to look for tag-SNPs in linkage disequilibrium (LD) with the inversion. Thus by examining the haplotypes of the surrounding regions, I can better estimate the inversion frequency using the publicly available HapMap SNP allele information. I obtained SNP genotypes and phased haplotypes from the HapMap Phase II project database for 180 CEU, 90 CHB, 91 JPT, and 180 YRI samples for all polymorphic inverted regions assayed in PCR experiments. I then searched for evidence of co-segregation by performing a correlation determination analysis using a minimum threshold $r^2 = 0.8$ between the inverted alleles and the phased SNP genotypes. The inversion frequencies in each population were then estimated

using the frequencies of the co-segregated haplotypes. Hence, for those regions where haplotype imputation was possible, I obtained a better estimate of allele frequencies. Naturally, I performed additional inversion PCR typing on samples predicted by SNPs to be inverted to verify my imputations.

Moreover, at these imputable regions, I further examined population differentiation and haplotype diversity. Population differentiation of individual variants was estimated by the statistic fixation index, F_{ST}. Finally, to examine haplotype diversity, I constructed haplotype networks based on phased HapMap SNPs surrounding each imputable inversion. The networks are built using a median-joining algorithm (Bandelt, et al., 1999) available in the SplitsTree software (Huson and Bryant, 2006).

IV.C Results

IV.C.1 Inversions in the human population

Accurate breakpoint information from HuRef variants offers an opportunity to genotype inversions in a larger number of individuals, to better understand their structure and frequency. In particular, I selected eight HuRef regions from 1.1 to 21.9 kb. Five of the eight were caused by non-homologous processes, and three by NAHR, and I looked for any difference in their structure and frequency. I designed PCR assays for targeted genotyping across DNA samples. The cohort consisted of a panel of 42 human samples (10 HapMap YRI, 10 HapMap CEU, 10 HapMap JPT, and 10 HapMap CHB, a phenotypically normal individual NA15510, and HuRef), three chimpanzee samples and one orangutan sample.

Three of the selected inversion regions (4q22.1, 1q31.3 and 16q24.1) were bi-allelic and polymorphic with unaltered breakpoints (Table IV. 2). They were formed by non-homologous processes. From primate sequences, two of the regions indicated the reference orientation to be the ancestral allele. For four of the eight regions, I was able to identify a tag SNP that is in LD ($r^2 \geq 0.8$) with the inversion, and I obtained a more accurate estimate of the inversion allele frequency using these SNPs as proxies (Table IV. 3). From the imputation results, I calculated the level of genetic differentiation, and found that three of the four loci showed similar allelic frequency across populations.

Table IV. 2. Summary of inversion genotyping experiments.

Locus	Coordinate	Size (bp)	Methods	Mechanism	Allele	European	Chinese	Japanese	Yoruban	Ancestral	Imputation	Tag SNP*	F_{st}**
3q26.1	chr3:164,008,436-164,030,337	21,902	Mate-pair	NH	Multi-allelic	-	-	-	-	-	No	-	-
Xp11.3	chrX:46,695,748-46,715,632	19,885	Assembly comparison	NAHR	Inversion	19	14	14	16	Inversion	No	-	-
					Reference	0	0	0	0				
7q11.22	chr7:70,058,905-70,076,823	17,919	Mate-pair	NH	Inv-del	14	13	15	3	Reference	Yes	rs1525303	0.38
					Reference	10	7	5	17				
16q23.1	chr16:73,797,599-73,814,159	16,561	Assembly comparison	NAHR	Inversion	15	19	20	20	Inversion	No	-	-
					Reference	5	0	0	0				
			-	-	Deletion	4	1	0	0				
4q22.1	chr4:89,066,188-89,077,724	11,537	Mate-pair	NH	Inversion	14	16	17	18	Inversion	Yes	rs1477602	0.08
					Reference	10	4	3	2				
1q31.3	chr1:196,023,411-196,024,609	1,199	Mate-pair	NH	Inversion	8	2	2	0	Inversion	Yes	rs1627999	0.17
					Reference	16	18	18	20				
6q27	chr6:168,835,529-168,836,601	1,073	Assembly comparison	NAHR	Inversion	24	20	20	20	Inversion	No	-	-
					Reference	0	0	0	0				
16q24.1	chr16:83,746,237-83,747,302	1,066	Assembly comparison	NH	Inversion	12	10	11	15	Reference	Yes	rs9933231	0.03
					Reference	12	10	9	5				

*A tag SNP must have an r-square value of at least 0.8

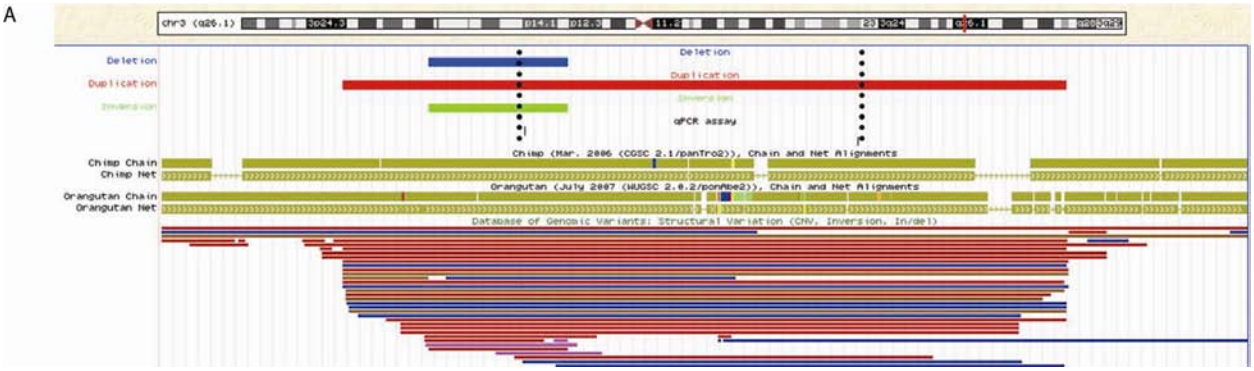
** F_{st} value is computed based on the frequency of a tag SNP

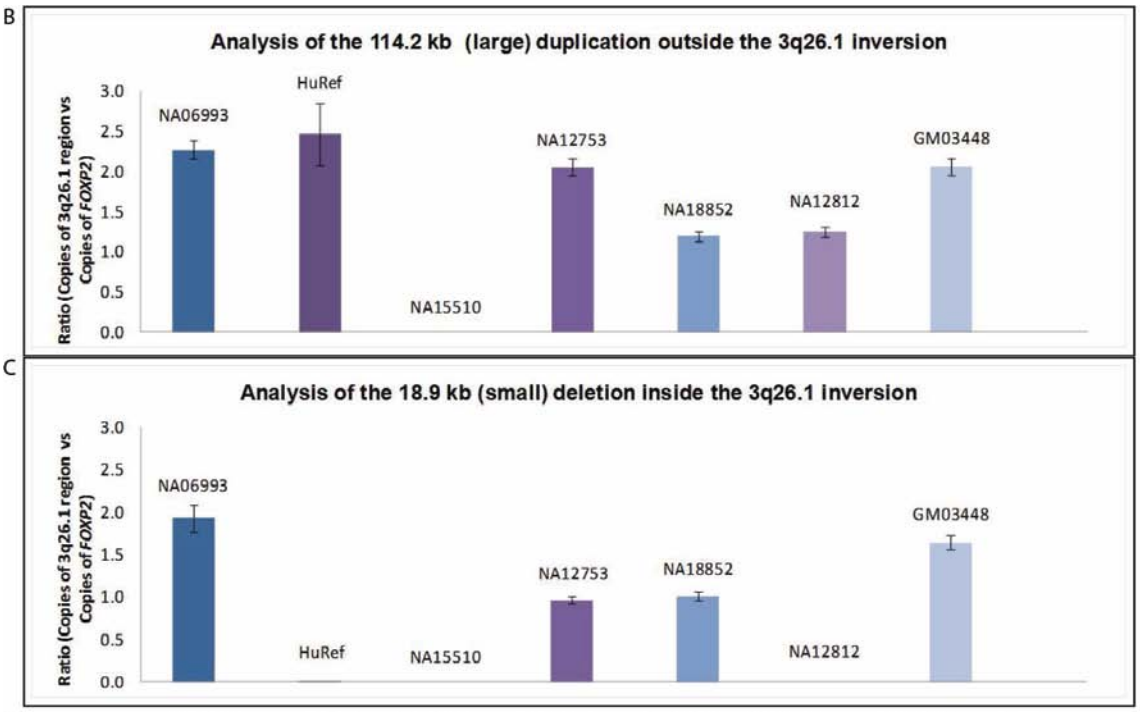
Table IV. 3. Inversion allele frequency as estimated by SNP-imputation.

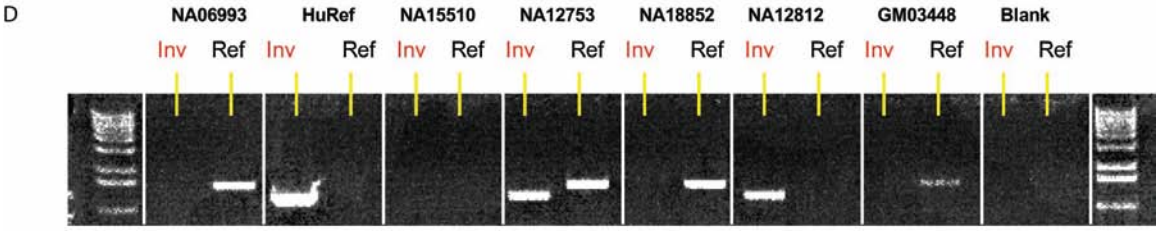
Locus	Tag SNP*	SNP allele	Inversion coordinate	Size (bp)	Methods	Allele	European	Chinese	Japanese	Yoruban
7q11.22	rs1525303	A	chr7:70,058,905-70,076,823	17,919	Mate-pair	Inv-del	83	65	59	28
		T				Reference	37	25	27	90
4q22.1	rs1477602	A	chr4:89,066,188-89,077,724	11,537	Mate-pair	Inversion	81	66	68	98
		G				Reference	39	24	22	16
1q31.3	rs1627999	G	chr1:196,023,411-196,024,609	1,199	Mate-pair	Inversion	24	13	10	1
		A				Reference	96	77	80	115
16q24.1	rs9933231	T	chr16:83,746,237-83,747,302	1,066	Assembly comparison	Inversion	81	51	51	77
		A				Reference	37	39	39	37

IV.C.2 Complex inversion structures

In Chapter III, I reported one 114.2 kb duplication, one 18.9 kb deletion, and one 21.9 kb inversion in HuRef in a highly complex region in 3q26.1, where many studies have also detected numerous CNVs and inversions (Figure IV. 2 A). Specifically, the duplication was detected by NimbleGen 42M array CGH, but the deletions and inversions were independently detected by mate-pair mapping (Pang, et al., 2010). Therefore for my current study, besides the inversion assay, I also designed two qPCR assays (one inside and one outside of the inverted locus) to genotype all three events. The outside qPCR assay aimed to target the 114.2 duplication, whereas the internal assay targeted the 18.9 kb deletion. Surprisingly, I noticed that 50% of individuals have a large deletion in place of the 114.2 kb duplication (Figure IV. 2 B). This variant is polymorphic and harbors different copy number states. Moreover, among those individuals with the 21.9 kb inversion, all have the 18.9 kb deletion embedded in the inverted area. I believe that the inversion and deletion may have arisen concurrently (Figure IV. 2 C and D). From these observations, I hypothesize that this region is multi-allelic harboring multiple polymorphisms. Also, since there is no segmental duplication in the region, I believe that replication-based mechanisms such as FoSTeS or MMBIR could be responsible for the observed complexity.







Samples	NA06993	HuRef	NA15510	NA12753	NA18852	NA12812	GM03448
Allele1	LDup	LDup-Inv-SDel	LDel	LDup	Ref	Inv-SDel	LDup
Allele2	LDup	LDup-Inv-SDel	LDel	LDup-Inv-SDel	Ref	Inv-SDel	LDup

Figure IV. 2. Complexity at the 3q26.1 region.

(A) The top three tracks represent variation detected in the HuRef genome. The HuRef inversion of interest is shown in the green track, a deletion in blue and a large duplication in red. Furthermore, notice all the variants discovered in pervious studies as shown in the DGV track. The vertical dotted lines represent locations targeted by qPCR assays. (B) QPCR targeting of the 114.2 kb duplication outside the 3q26.1 inversion. (C) QPCR analysis of the 18.9 kb deletion inside the 3q26.1 inversion. (D) PCR targeting the 3q26.1 inversion. Notice that inversion and small 18.9 deletion always occur together. Note that besides the chimpanzee sample GM03448, all others are human samples. The inferred genotype for each sample is listed below the gel image.

Another complex example is at 7q11.22, where there is co-occurrence of a 12.7 kb inversion and a 5.2 kb deletion – a situation supported by a previous study comparing the chimpanzee and the human reference assembly (Feuk, et al., 2005). A non-homologous process formed this inversion. By direct genotyping and SNP-imputation, I discovered that the inverted allele became the major allele in Europeans and Asians, but remained as a minor allele in Africans (Figure IV. 3 A). I observed more haplotypes with the reference genome orientation (Figure IV. 3 B to E), which is also the orientation found in chimpanzee, suggesting that the reference assembly contains the ancestral allele (Table IV. 2). Although I found no genes or regulatory elements in the locus, there was evidence of population differentiation, with F_{st} = 0.38, which indicates 38% of allele frequency variance is found between different populations – much higher than the 10% value typically found between population groups (Conrad, et al., 2010b; Durbin, et al., 2010). In the absence of any functional elements in the locus, I postulate that founder effect in the Eurasian ancestral population and genetic drift most likely explain the allele frequency difference observed between Africans and Eurasians.

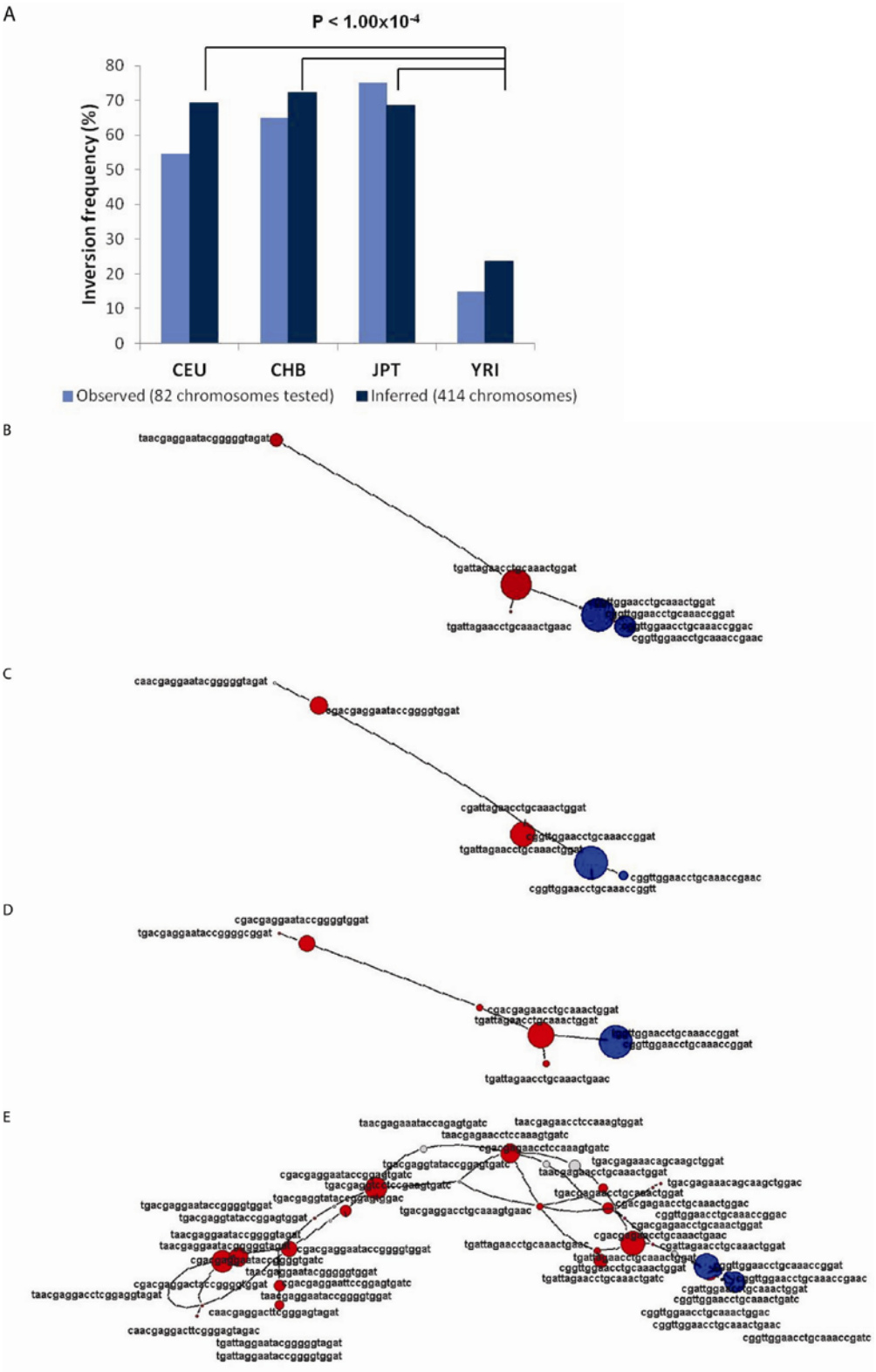


Figure IV. 3. 7q11.22 inversion allele distribution among four HapMap III populations.

(A) Observed frequency from PCR genotyping and imputed frequency of the 7q11.22 inversion. (B to E) Haplotype network graphs. The size of nodes represents haplotype frequency in the HapMap 2 cohorts, while the clades represent the amount of nucleotide substitution difference between adjacent nodes. Blue and red nodes correspond to haplotypes with the inverted and reference allele, respectively. Grey nodes mean that the orientation cannot be determined. (B) CEU samples. (C) CHB samples. (D) JPT samples. (E) YRI samples.

In HuRef, I found a 16.6 kb inversion at 16q23.1 that disrupts two genes – *CTRB1* and *CTRB2* – such that two genes with exchanged exon-1were potentially created (Figure IV. 4). Both *CTRB1* and *CTRB2* are members of the chymotrypsinogen B precursor. The two genes share an overall 97% DNA sequence identity, yet their first exons, which are protein-coding, are only 82% similar. The NAHR-associated inversion is the ancestral allele, and is highly prevalent in the population: 37 (out of 42) individuals were homozygous for the inversion, and five were heterozygous (Table IV. 2). Interestingly, all five individuals were of European origin. Public GenBank RNA databases showed five records with the exchanged transcript sequence. Specifically, entries M24400.1, BC005385.1, and BT007356.1 showed exon-1 of *CTRB2* followed by exons of *CTRB1*, whereas entries BC073145, AK131056 had exon-1 of *CTRB1* followed by exons of *CTRB2*. In four Europeans and one Chinese, I identified an adjacent 585 bp deletion that overlapped the entire 134 bp exon-6 of *CTRB2*, thus creating an out-of-frame transcript product. This deletion was not observed in the HuRef sample. Also, the deletion was found only on chromosomes with the 16.6 kb inversion. Considering that the primate samples were homozygous for the inversion, I postulate that the deletion was a derived allele that arose on an inverted haplotype. Finally, I found a *CTRB2* transcript entry (AW584011.1) in the GenBank EST database that does not have an exon-6, which would correspond to the deletion allele found in this study. The corresponding results between genomic variation data and transcriptomic data highlight the importance of correlating both data types to delineate the structure and function of the human genome (McPherson, et al., 2012).

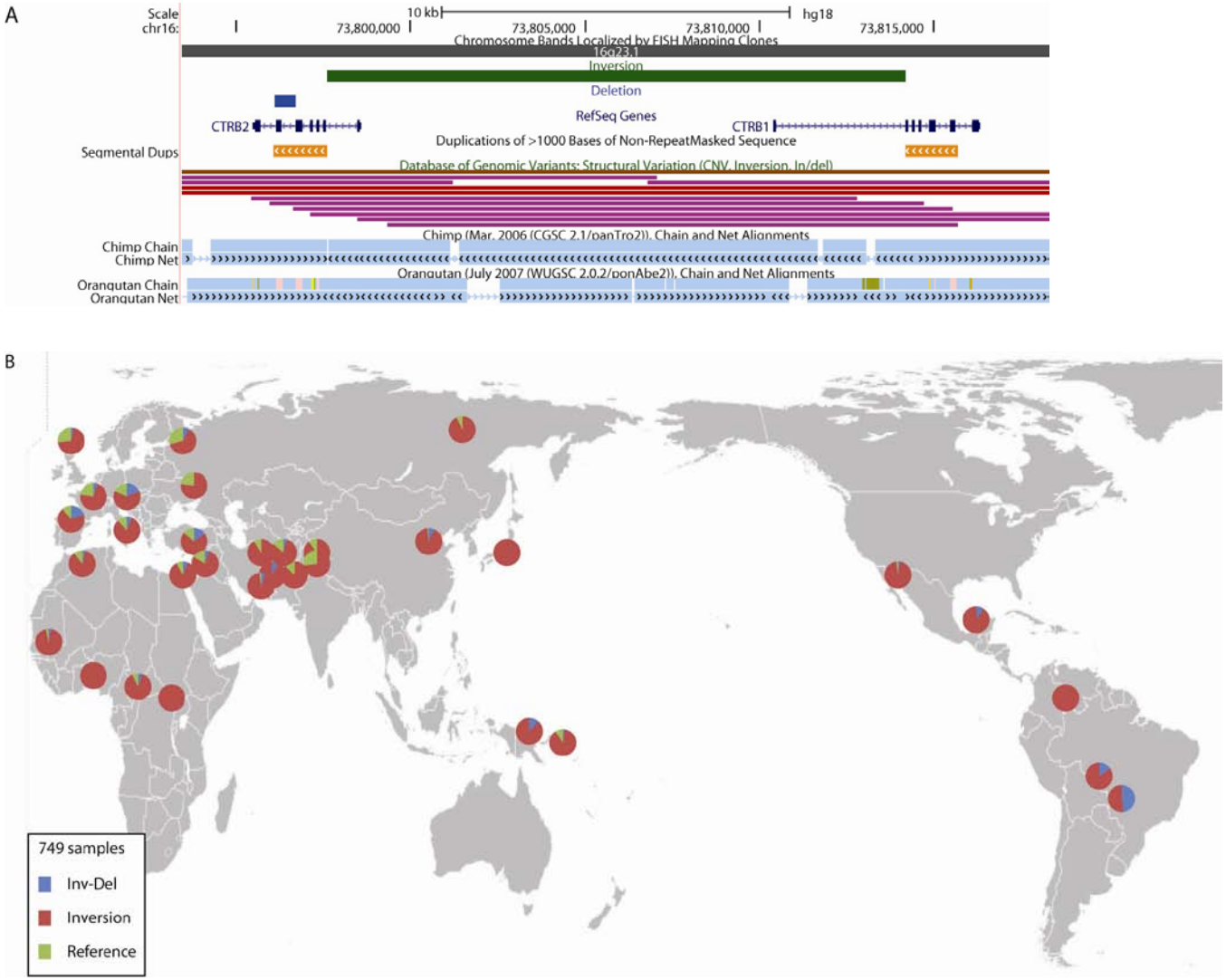


Figure IV. 4. 16q23.1 inversion and the associated deletion.

(A) A genome browser showing the positions of the inversion (top green track) and deletion (second blue track), and the impacted genes CTRB1 and CTRB2 are shown. (B) Haplotype frequency showing HGDP-CEPH populations where at least 10 samples were genotyped. A total of 871 samples have been genotyped, whereas 749 are displayed here.

In light of the interesting finding at this 16q23.1 region, additional samples were tested with the same assay to better estimate the allele frequency. A total of 871 HGDP-CEPH samples from 57 populations were genotyped (Figure IV. 4 B). Consistent with the HapMap sample results, the inversion was the major allele. Moreover, the deletion was only observed in the inverted haplotype. The inversion-deletion haplotype is most prevalent in Surui (47.2 %), French Basque (20.5 %), North Italian (17.9 %) and Druze (15.4 %), and lowest with a frequency of zero in Yoruba, Yakut, Sindhi, and Mbuti Pygmies. Eleven (out of 871 samples) were homozygous for the inversion and deletion. The F_{st} value was 0.53, so there was evidence of population differentiation in haplotype frequency.

IV.C.3 Dynamic regions in the human reference assembly

Eight inversions were genotyped, five of which are non-homologous events and the other three are NAHR events. All of the bi-allelic and imputable variants were formed by non-homologous events, but the NAHR-derived loci were more complex. As mentioned above, the inversion at 16q23.1 was associated with an additional deletion, creating fused and non-functional genes. The other two showed potential reference assembly errors (Table IV. 2). Particularly, a 19.9 kb inversion at Xp11.3 was flanked by two Alu elements. At this region, NCBI Build 36 reference and subsequent Build 37 both showed that the supposed inversion is located at the edge of an assembly clone Z83822.2; however, neither the genotyping results nor the chimpanzee assembly supported the human reference orientation. The GenBank record was updated on July 10, 2011, and the clone was trimmed such that it no longer covers the 19.9 kb region of interest. Instead, the region is now represented by the neighboring clone AL627143.15, and its orientation is now concordant with the genotyping results (Figure IV. 5).

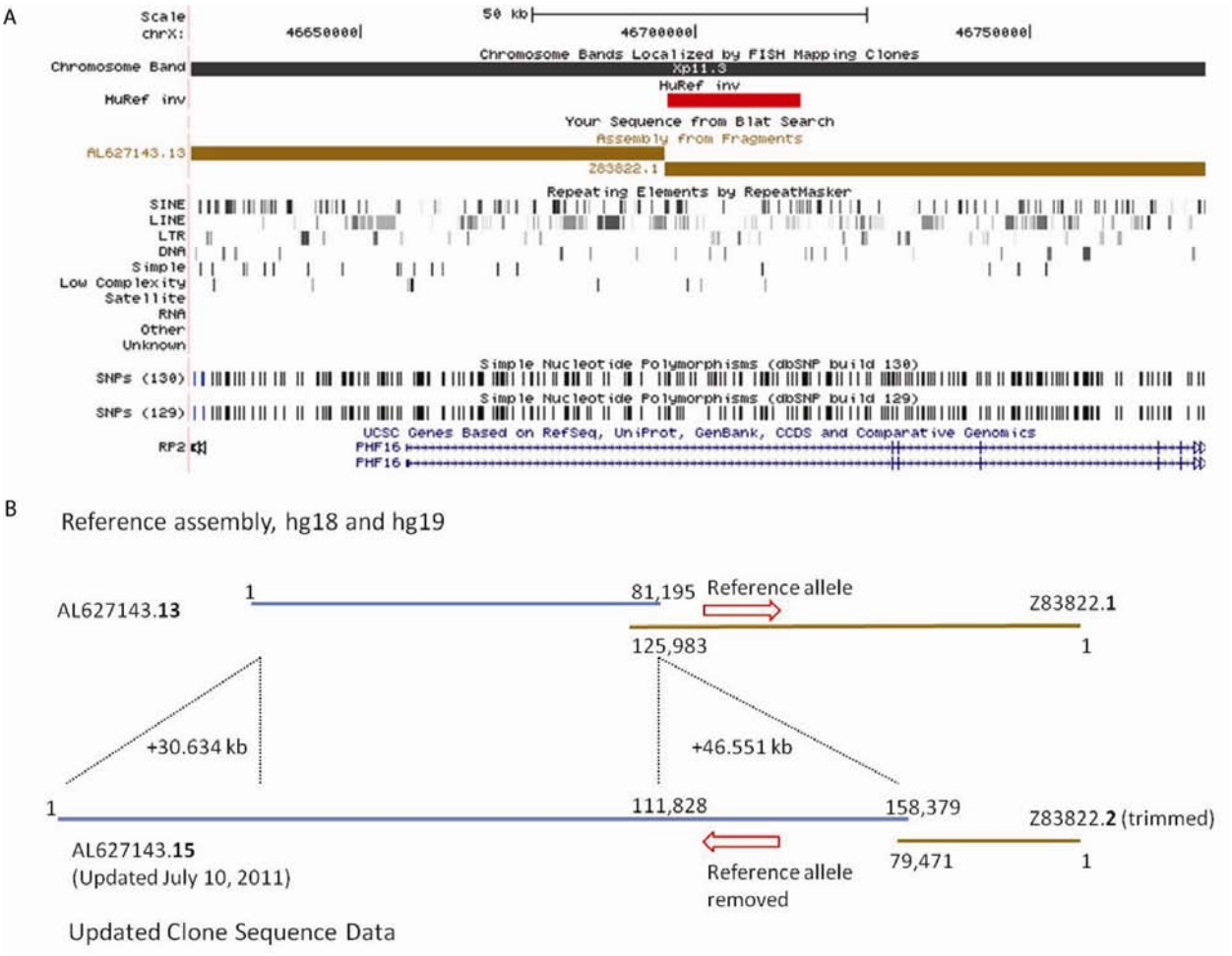


Figure IV. 5. The Xp11.3 region.
(A) UCSC genome browser screenshot of the region. The red track is the HuRef inversion of interest. Notice that it resides at the edge of Z83822.1 assembly clone. (B) Schematic of the change in the reference assembly and the switch in sequence direction. As seen in Table 1 in the main text, there is no sample having the reference genotype, thus suggesting a putative reference assembly error. In both NCBI Build 36 and 37, the inversion of interest resides in the Z83822.1 clone, but in the update on July 10th, 2011, the region is covered by the neighboring clone AL627143.15. The AL627143.15 has been extended in both direction, and in doing so, the orientation at the region of interest has been flipped, thus removing the original reference orientation.

Next, I compared all of the HuRef inversions with the DGV (Iafrate, et al., 2004; Zhang, et al., 2006). I identified 15 loci where the majority of previous studies had unanimously called inversions. Again, most of them (11 out of 15) were NAHR-derived variants. I postulate that the reference genome orientation of these loci represents either a minor allele or is incorrect. Particularly, the reference assembly clones in five of these regions had been modified, and their orientation reversed from Build 36 to 37, thus indicating potential errors in the Build 36 (Table IV. 4). To further investigate the number of inverted regions which have been highlighted as problematic, and may undergo additional modifications in upcoming assemblies, I compared the inversion dataset to the list of regions targeted by The Genome Reference Consortium for manual review and additional sequencing (Church, et al., 2011). I found that 49 of 166 regions coincide, thus indicating that these regions may not yet be fully resolved and require additional experimentation to determine the accurate structure. An alternate explanation is that there are indeed two (or more) alleles in humans, but the specific allele represented has changed over time. These changes exemplify the dynamic nature of the reference assembly.

Table IV. 4. List of HuRef inverted regions which are also discovered by previous inversion studies as listed in the DGV.

Co-ordinates*	Detection Method	Mechanism	Size (bp)	Ahn, et al., 2009	McKernan, et al., 2009	Tuzun, et al., 2005	Kidd, et al., 2008	Kidd, et al., 2010	Korbel, et al., 2007	Chimpanzee Assembly	Orangutan Assembly	Human reference assembly change
chr21:26296022..26296571	mate pair	blunt_end	550	y	y	n	n	n	n	y	y	no change, same clone same version used from B35 to B37
chr16:83746237..83747302	assembly comparison	blunt_end	1,066	y	y	n	n	n	y	y	y	no change, same clone same version used from B35 to B37
chr6:168835529..168836601	assembly comparison	nahr	1,073	n	y	n	y	n	y	y	y	no change, same clone same version used from B35 to B37
chr12:12436020..12437892	mate pair	nahr	1,873	y	y	n	n	n	y	y	y	no change, same clone same version used from B35 to B37
chr7:106846108..106850529	mate pair	nahr	4,422	n	y	n	y	n	y	y	y	no change, same clone same version used from B35 to B37
chr16:54399633..54405700	mate pair	imprecise breakpoint	6,068	y	y	y	y	y	n	y	n	no change, same clone same version used from B35 to B37
chr3:50900409..50910036	assembly comparison	nahr	9,628	n	n	y	y	n	y	y	y	same clones used, different portion of clone AC099047.2 used/trimmed (-80kb), clone AC131013.2 extended (+74kb), removed

												reference orientation.
chr9:125780830..125791433	mate pair	imprecise breakpoint	10,604	n	n	y	y	y	y	y	y	no change, same clone same version used from B35 to B37
chr12:79370385..79381831	assembly comparison	nahr	11,447	n	n	n	y	y	y	y	y	no change, same clone same version used from B35 to B37
chr12:85764376..85777047	assembly comparison	nahr	12,672	n	n	n	y	y	y	y	y	no change, same clone same version used from B35 to B37
chr1:2475133..2489144	assembly comparison	nahr	14,012	n	y	y	y	n	y	y	y	clone AL139246.21 updated from AL139246.20 in hg19, removed reference orientation allele
chr2:234136102..234151235	assembly comparison	nahr	15,134	n	n	n	y	y	y	y	y	clone AC019072.78 updated from AC019072.7 in hg19, region from ~96 kb to 114kb flipped, removed reference orientation allele
chr2:234135688..234151644	mate pair	nahr	15,957	n	n	n	y	y	y	y	y	clone AC019072.78 updated from AC019072.7 in hg19, region from ~96 kb to 114kb flipped,

												removed reference orientation allele
chrX:46695748..46715632	assembly comparison	nahr	19,885	n	y	y	y	n	y	y	y	clone AL627143.15 updated from AL627143.13 on July 10, 2011, extended 46.6 kb and removed the reference orientation allele in the region of interest.
chr16:54354976..54423120	mate pair	nahr	68,145	y	y	y	y	y	n	n	n	no change, same clone same version used from B35 to B37

*Dark green box shows evidence of inversion existing in the dataset, while red box shows no evidence of inversion .
I expect that inversion based on short insert mapping (Ahn and McKernan) can capture small size inversion,
while those relying on large insert mapping (Tuzun, Kidd 2008 and 2010, Korbel) to be able to identify large inversion.

IV.D Discussion

The availability of breakpoint-resolution allows for characterization in the general population of inversion polymorphisms, which have been underrepresented in most genomic studies. While three out of eight inversions were bi-allelic, had simple breakpoint junctions, and were tag-able by neighbouring SNPs, the rest were either multi-allelic, contained complex rearrangements, or were potentially reference assembly errors. From the results, I saw that small, presumably benign, polymorphic inversions could be complex, involved concurrent gain or loss of additional DNA sequences, and were similar in structure to larger inversions that had been associated with disease (Antonacci, et al., 2009; Chiang, et al., 2012; Kloosterman, et al., 2011; Osborne, et al., 2001; Stephens, et al., 2011).

I believe that my assay design is most successful in typing bi-allelic, imputable inversions with simple junction structures. The ascertainment of complex, recurrent and NAHR-related variants will require a combination of longer sequence lengths, targeted local assembly, and long-range haplotyping (Bansal and Bafna, 2008; Fan, et al., 2011; Khaja, et al., 2006; Kitzman, et al., 2011; Levy, et al., 2007; Scherer, et al., 2003). Moreover, ultramicro-inversions on the other end of the size spectrum are largely understudied (Hara and Imanishi, 2011). These variants can be detected within sequence reads, and are likely to have blunt-end boundaries, similar to most of the annotated indels discussed in Chapter III.

I observed an inversion at 16q23.1 where an inversion and a deletion may potentially impact the function of the *CTRB1* and *CTRB2* genes. There was evidence of population differentiation in haplotype frequency. Moreover, despite the fact that both gene products are homologous, and are expressed in pancreatic islet cells in the kidney, there may be differences in expression pattern. According to ENCODE Project Consortium (Myers, 2011), there is a denser cluster of transcription factor binding sites and promoter-associated histone marks upstream of the exon -1 of *CTRB1* than *CTRB2* (Myers, 2011), and so an exchange of the exon-1 by the 16.6 kb inversion may alter the expression patterns in addition to the creation of hybrid protein structures. In addition, the frame-shift deletion would disrupt the trypsin-like serine protease domain. Chymotrypsinogen B is the precursor to the digestive enzyme chymotrypsin, whose function is to cleave aromatic amino acids such as

phenylalanine, tyrosine and tryptophan. The observed difference in haplotype frequency across population may be the result of adaptation to different diet. Further studies will be required to elucidate the functional impact of the variations characterized here at the DNA level.

In conclusion, with precise breakpoint information, I annotated a subset of the HuRef inversions in the human population, and identified the inverted allele frequencies. I identified inversion alleles that exhibit population differentiation, and impact genes. Most importantly, I discovered that inversions can be associated with other rearrangements, creating more complex structures. These structures may even be challenging to the reference assembly. This study offers additional insights into the origin and complexity of the often understudied submicroscopic inversions.

CHAPTER V: SUMMARY AND FUTURE DIRECTIONS

V.A Summary and future directions

Our concept of what is variable in the genome has changed dramatically over the past half century: from rare chromosomal rearrangements, to single base polymorphisms, to various forms of submicroscopic structural variation. The field of variation study changes from investigating single locus to whole genome, and with ever improving accuracy and sensitivity. We are now able to study the full complements of variation in an entire population cohort. In recognition of the progress achieved in variation research, "Human Genetic Variation" is considered to be the breakthrough of the year by the Science magazine in 2007 (Pennisi, 2007). I showed that over 1 % of the genome is variable, and the majority of which are due to structural variants. The work presented in this thesis contributes to our understanding of variation by defining the amount of variation content between any two genomes, highlighting strengths and limitations of structural variation discovery methods, quantifying the different structural variation-formation mechanisms, and examining the structure and frequency of inversions.

V.B Remaining challenges

Technology has been instrumental in driving discovery. Presently, NGS holds great promise in impacting biomedical research. It can produce an unprecedented amount of sequence information at a low-cost and high throughput fashion. However, there are still some shortcomings in current technologies.

V.B.1 Gap in variant discovery

Current and future genome sequencing experiments using NGS technologies will become an increasingly common and inexpensive approach to discover variation within personal genome sequences. However, the improved speed and decreased cost come with a number of challenges, and most notably a reduction in resolution to detect all types and classes of genetic variation (Pang, et al., 2010). While NGS detection of single nucleotide and very small indels seems sufficient (Lam, et al., 2012), the short read lengths of NGS would limit the detection of larger and more complex genetic variants. The HuRef variation set described in this thesis (termed the HuRef Standard in this section) can act as a baseline to compare

variation data generated by NGS, and to investigate the completeness and accuracy of calls. In other words, if one were to sequence the HuRef genome by a NGS platform, one can directly examine the sensitivity and specificity of NGS data.

The HuRef genome has also been sequenced by Complete Genomics (CG) (Drmanac, et al., 2010). The CG platform was chosen because of its standardized sequencing process and analysis pipeline, its wide spread use, and its robustness in variation-detection performance (Figure V. 1, Table V. 1). With paired-end sequencing of inserts approximately 400 bp in length, an average depth of coverage of 63.5X was achieved in sequencing the HuRef genome by CG. The CG variant calls were detected primarily by three approaches: split-read, paired-end and read depth. Note that paired-end mapping is the same as mate-pair mapping, except that the insert fragment of a paired-end library (200-400bp) is smaller than a mate-pair library (a few kilobases).

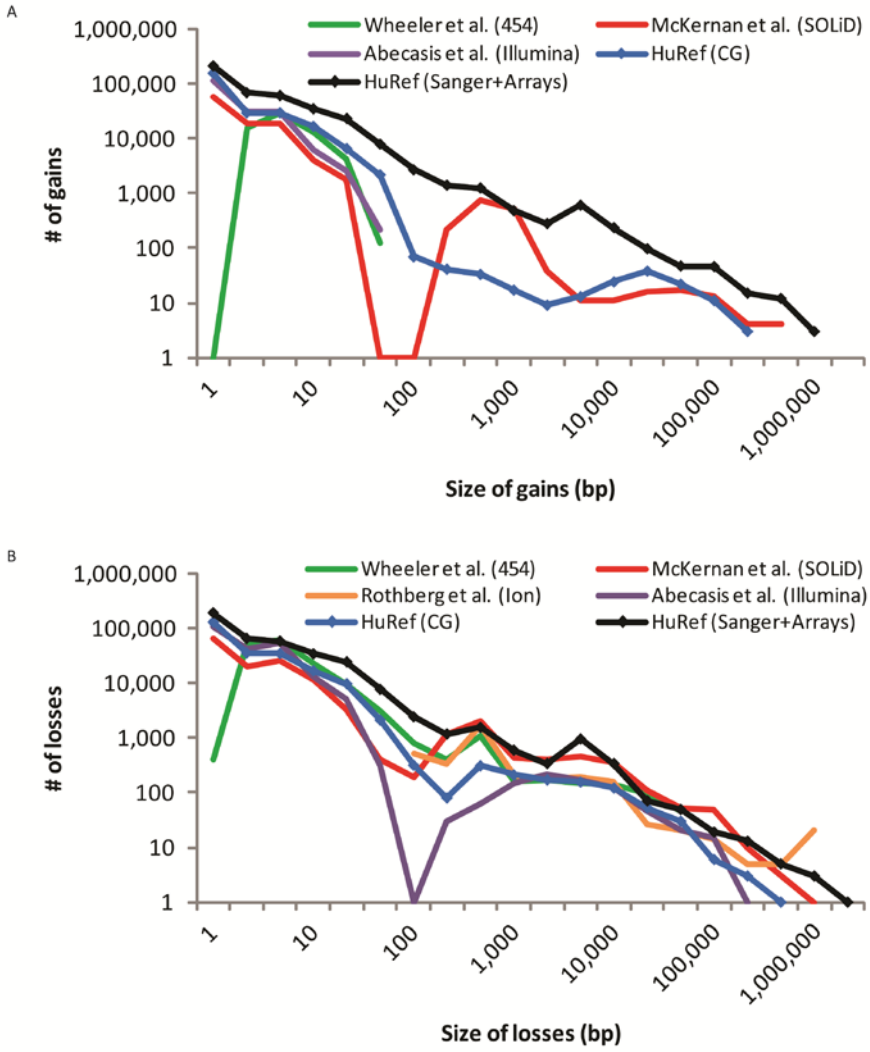


Figure V. 1. The size distributions of reported DNA gains and losses in published personal genome sequencing studies.

These diagrams show the relative uniformity of CG variants across the size spectrum. In Wheeler et al., insertions identified by intra-read alignment would be limited by the size of the 454 sequencing reads; hence, large insertions beyond the read length were not detected (Wheeler, et al., 2008). McKernan et al. used SOLiD and microarrays to detect variation in Yoruba individual NA18507 (McKernan, et al., 2009). They detected small variants based on split-reads and large ones based on mate-pair and microarrays, but failed to find medium size gains. Rothberg et al. performed whole genome sequencing using the Ion Torrent technology, but only reported deletions at least 50 bp in size (Rothberg, et al., 2011). Mainly relying on Illumina, Abecasis and colleagues detected variation in the sample NA18507 using a multitude of calling algorithms (Abecasis, et al., 2012). However, for large variation, only deletions were reported. From these size distributions, CG yielded the most consistent calling pattern across the size spectrum when compared with other NGS technologies.

Table V. 1. Summary of variation results in some personal genomes

Sample	Pop.	Platform	Cov.	Gain/loss			Reference	
				#	Min size (bp)	Max size (kb)	Study	PMID
Venter (HuRef)	Caucasian	ABI3730xl; microarrays	7.5	796,079	1	82.7	Levy et al., 2007; Pang et al., 2010	17803354;20482838
Watson	Caucasian	454	7.4	222,718	2	38.9	Wheeler et al., 2008	18421352
NA18507	Yoruba	SOLiD	17.9	232,124	1	97	McKernan et al., 2009	19546169
Moore	Caucasian	Ion Torrent	10.6	3,391	50	982.8	Rothberg, et al., 2011	21776081
NA18507	Yoruba	Illumina	~30	405,741	1	100.5	Abecasis et al., 2012	23128226
Venter (HuRef)	Caucasian	Complete Genomics	63.5	471,770	1	16,797	Current chapter	

First, by examining the HuRef CG and HuRef Standard variation profile, one would notice that short read sequencing had challenges in detecting variants of certain size ranges. In total, there were 241,033 CG gains and 230,737 losses in the HuRef genome, which accounts for a portion of the HuRef Standards' 408,403 gains and 383,470 losses (Table V. 2). Unlike the uniform negative slope of the size distribution of variants annotated in the Sanger-based assembly of the HuRef Standard, there were notable drops in sensitivity in the CG set, particularly for gains in the paired-end detection range (Figure V. 2, Figure V. 3). As has been acknowledged (in CG Support & Community webpage), CG's junction detection approach has difficulty in calling variants at high identity repeats, and calling insertion sequences not in the NCBI reference genome. Also, in order to substantiate that the CG profile is indeed missing variants, not simply overcalling in the HuRef Standard set, one can compare the HuRef Standard variants with published studies. For example, one could compile 3,751,689 non-redundant variants from 18 published studies that have used multiple variant-detection methods: NGS, Sanger read-trace, Sanger fosmid-end mapping, and microarrays (Table V. 3) (Abecasis, et al., 2012; Alkan, et al., 2009; Altshuler, et al., 2010; Conrad, et al., 2010b; Durbin, et al., 2010; Itsara, et al., 2010; Jakobsson, et al., 2008; Ju, et al., 2010; Kidd, et al., 2008; Kidd, et al., 2010a; Kidd, et al., 2010b; McCarroll, et al., 2008; Mills, et al., 2011a; Perry, et al., 2008; Pinto, et al., 2011; Teague, et al., 2010; Tong, et al., 2010; Wheeler, et al., 2008). Then after cross-examining the HuRef Standard with this reference set, one would notice that the size distribution curves representing the HuRef Standard variants also detected in published studies would still be consistently at or above the overall HuRef CG curves across the entire size spectrum (Figure V. 4). Evidently, variants were missing the HuRef CG profile.

Table V. 2. Gains and losses detected in the HuRef genome by different methods

Detection strategy	Type	#	Min size (bp)	Median size (bp)	Max size (bp)	Total size (bp)
Sanger assembly comparison	*Hom ins*	*275,417*	*1*	*2*	*82,711*	*3,110,678*
	Hom del	*283,738*	*1*	*2*	*18,484*	*2,813,857*
	Het ins	*128,084*	*1*	*1*	*321*	*299,562*
	Het del	*92,564*	*1*	*1*	*316*	*220,051*
Sanger split-read	*Ins*	*3,747*	*11*	*18*	*414*	*125,549*
	Del	*5,577*	*11*	*16*	*111,714*	*1,141,842*
Sanger mate-pair	*Ins*	*656*	*346*	*3,566*	*28,344*	*3,177,629*
	Del	*1,077*	*352*	*3,827*	*232,308*	*5,034,418*
Agilent 24M	*Dup*	*136*	*445*	*993*	*81,458*	*457,872*
	Del	*217*	*439*	*877*	*852,404*	*2,157,491*
NimbleGen 42M	*Dup*	*357*	*448*	*4,672*	*836,362*	*11,098,815*
	Del	*293*	*459*	*2,712*	*359,736*	*3,634,700*
Affymextrix 6.0	*Dup*	*7*	*14,485*	*42,798*	*640,474*	*1,519,885*
	Del	*4*	*10,176*	*48,721*	*123,797*	*231,415*
Non-redundant total	***Gains***	***408,403***	***1***	***1***	***836,362***	***19,789,990***
	Losses	***383,470***	***1***	***2***	***852,404***	***15,233,774***

Detection strategy	Type	#	Min size (bp)	Median size (bp)	Max size (bp)	Total size (bp)
CG split-read	Ins	240,813	1	1	63	584,548
	Del	229,676	1	1	187	654,829
CG paired-end*	Dup	116	49	242	94,707	280,983
	Del	956	236	868	16,797,153	19,400,558
CG read depth	Dup	104	1,307	14,001	160,001	2,448,564
	Del	105	2,001	12,001	110,001	1,822,961
Non-redundant total**	**Gains**	**241,033**	**1**	**1**	**160,001**	**3,314,095**
	Losses	**230,737**	**1**	**1**	**16,797,153**	**21,878,348**

italics: generated from the non-redundant set from Levy et al., 2007, and Pang et al., 2010, and then subsequently lifted over from Build 36 to Build 37

*The 16.8 Mb deletion detected by CG paired-end approach is likely an artifacts, as it has not been detected by karyotype (Levy et al., 2007). Also, it was found in all the other 79 samples sequenced in this study. The next largest call is 242,290 bp.

**Excluding the CG paired-end 16.8 Mb deletion, the next largest CG deletion would be 242,290 bp, and total size would be 5,081,195 bp.

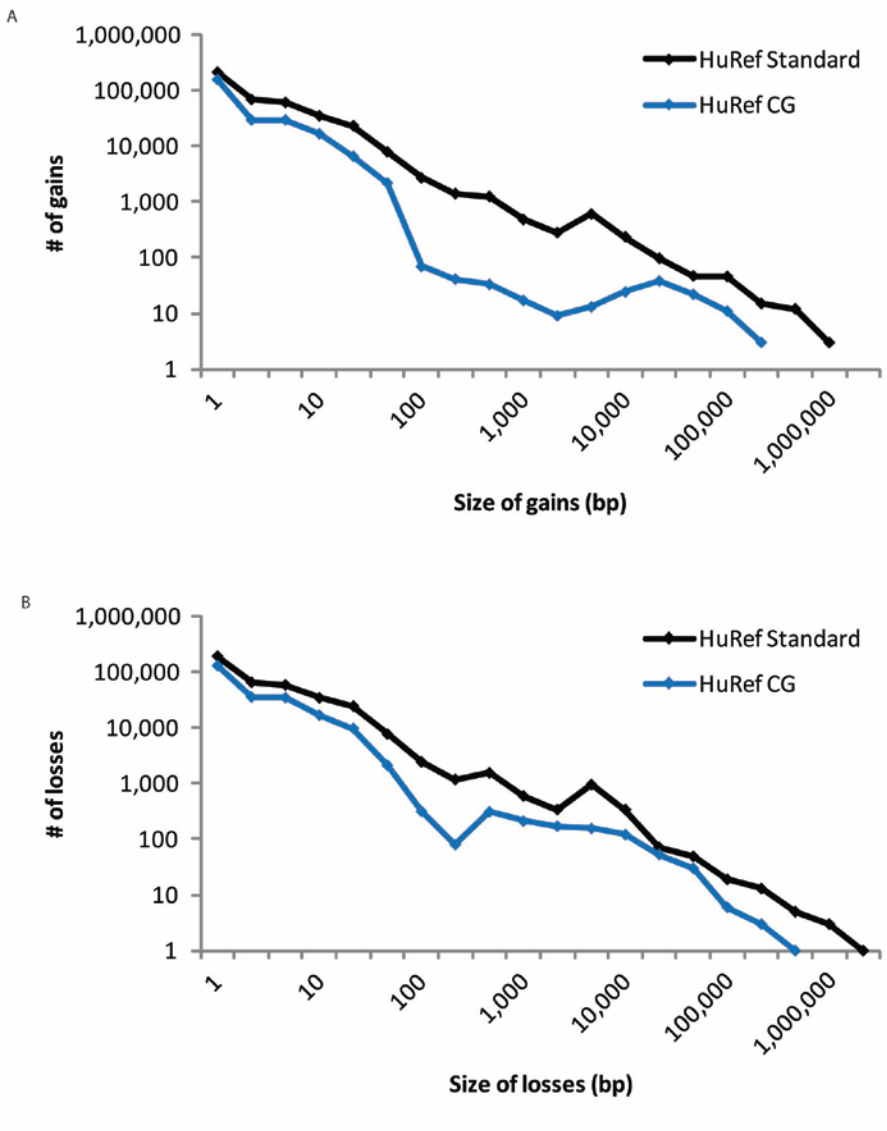


Figure V. 2. Size distribution of non redundant gains and losses detected in the HuRef sample.
(A) Gains. (B) Losses.

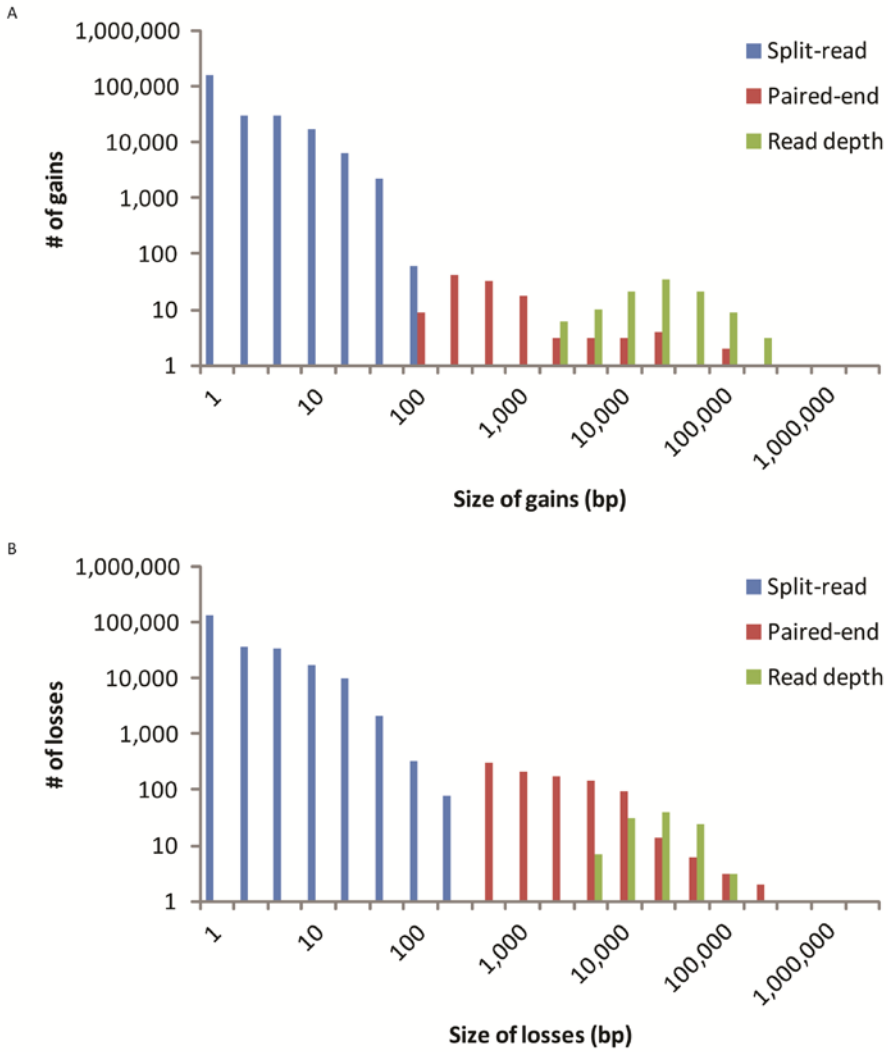


Figure V. 3. The size distributions of HuRef CG gains and losses detected by their discovery strategies.

Note that these two graphs (A for gains and B for losses) show all the calls detected by each approach, regardless of redundancy.

Table V. 3. Summary of variation results from published population studies.

Study	PudMed Id	Platform	Gains (#)	Losses (#)
Jakobsson et al., 2008	18288195	Genotyping array	9,102	4,626
Perry et al., 2008	18304495	Array CGH	26,730	8,578
Wheeler et al., 2008	18421352	NGS	284,346	156,770
Kidd et al., 2008	18451855	Sanger fosmid	15,597	3,961
McCarroll et al., 2008	18776908	Genotyping array	1,620	1,012
Itsara et al., 2009	19166990	Genotyping array	17,699	9,939
Alkan et al., 2009	19718026	NGS	1,154	42
Conrad et al., 2009	19812545	Array CGH/Genotyping array	77,762	25,744
Kidd et al., 2010	20440878	Sanger fosmid	14,318	0
Teague et al., 2010	20534489	Optical mapping	8,639	2,117
Ju et al., 2010	20802225	Array CGH	1,574	1,010
Altshuler et al., 2010	20811451	Genotyping array	173,254	142,752
Tong et al., 2010	20822512	NGS	286,704	104,360
Durbin et al., 2010	20981092	NGS	2,244,804	2,045,714
Kidd et al., 2010	21111241	Sanger fosmid	1,469	627
Mills et al., 2011	21460062	Sanger trace	2,933,141	976,321
Pinto et al., 2011	21552272	Array CGH/Genotyping array	76,878	43,334
Abecasis et al., 2012	23128226	NGS	2,043,940	888,150
Non redundant total			**1,637,756**	**2,113,933**

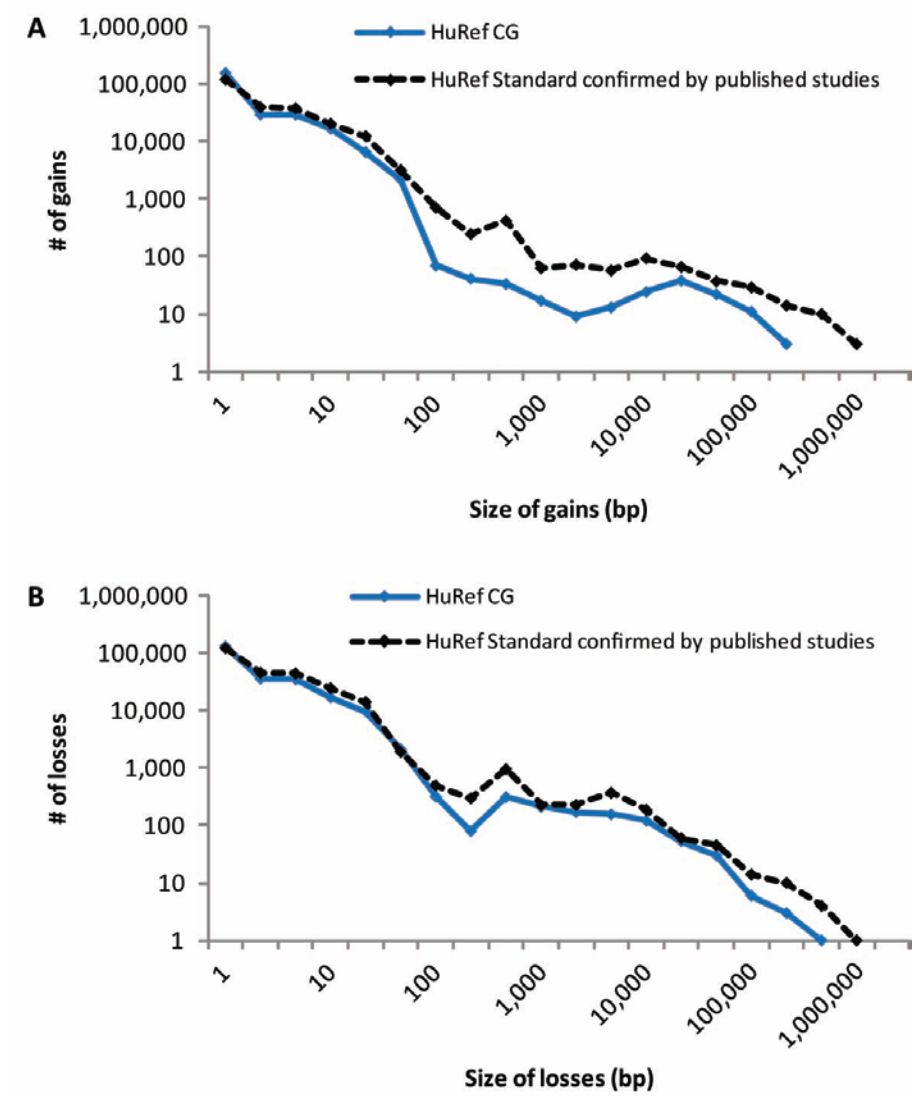


Figure V. 4. The size distribution of HuRef Standard variation that was confirmed by published studies.

The distributions for gains and losses are shown in plots (A) and (B), respectively. Note that the confirmed HuRef Standard size distributions were consistently equal to or above the HuRef CG ones.

Some of the missing gains and losses reside in DNA regions containing repeats. There can be notable reduction of calls in loci with retrotransposable repeats, tandem repeats and segmental duplications in the HuRef CG data with respect to the HuRef Standard (P-value < 2.2e-16) (Figure V. 5). These observations highlight the importance of having long reads and long inserts for alignment and variant-calling. As for centromeric and telomeric repeats, both Sanger sequencing and HTS have challenges at these regions, it is premature to evaluate their variant-calling performance.

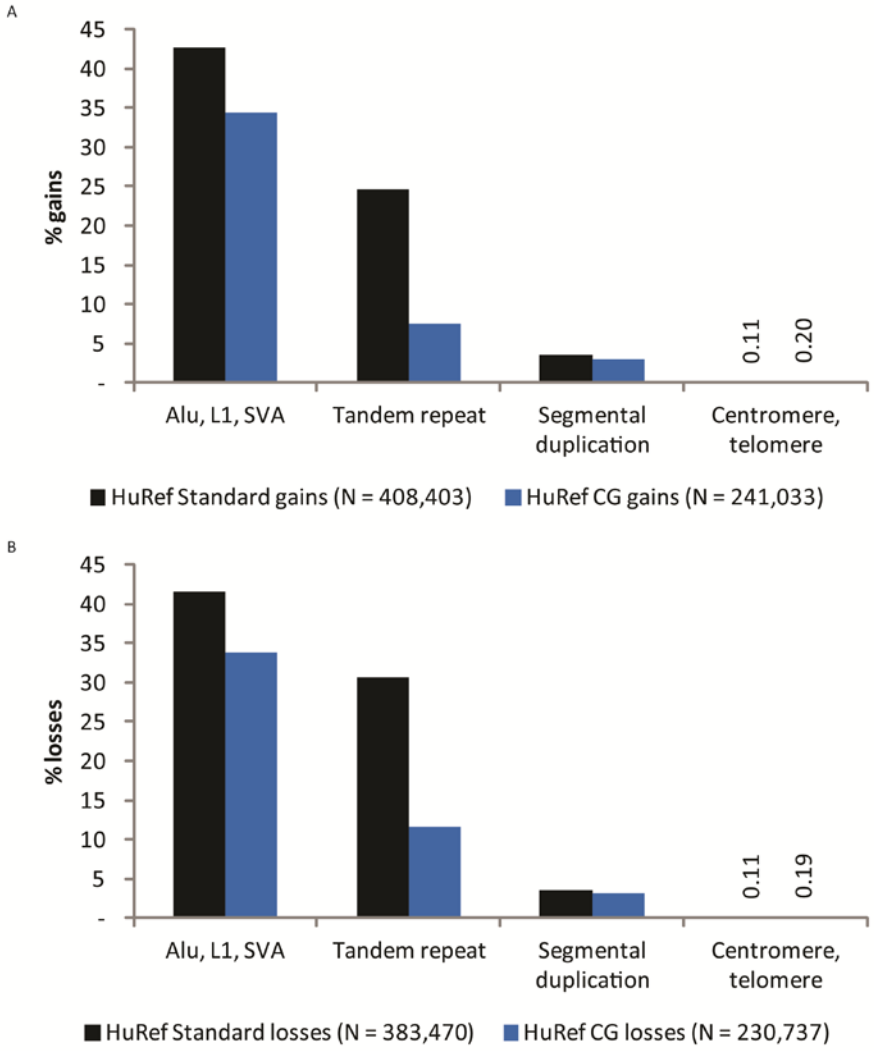


Figure V. 5. Proportion of HuRef Standard and HuRef CG gains and losses residing in repetitive regions.
(A) Gains. (B) Losses.

While there was a good overall concordance rate (64.2%) for the CG calls with the HuRef Standard, the specificity of gains would be lower than that of losses. About 59.1% (142,368/241,033) of gains and 69.5% (160,392/230,737) of losses called by CG were concordant (70% reciprocal size overlap) with the HuRef Standard (Figure V. 6).

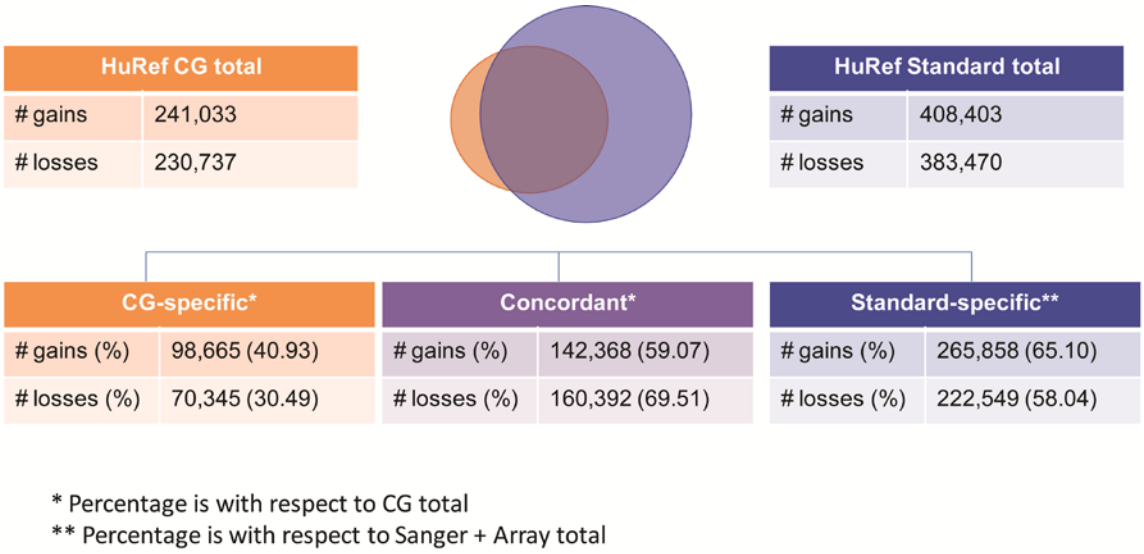


Figure V. 6. Overall concordant statistic between HuRef Standard and HuRef CG variation sets.

From comparison of HuRef CG and HuRef Standard, one can see that CG also has notable strengths. First, the HuRef CG loss size distribution was fairly uniform compared to the expected HuRef Standard (Figure V. 2). Second, CG was highly precise in determining variant size, with the exception of overcalling by the read-depth approach (Figure V. 7). Decreasing the binning-size together with increasing sequencing coverage can reduce the overestimation.

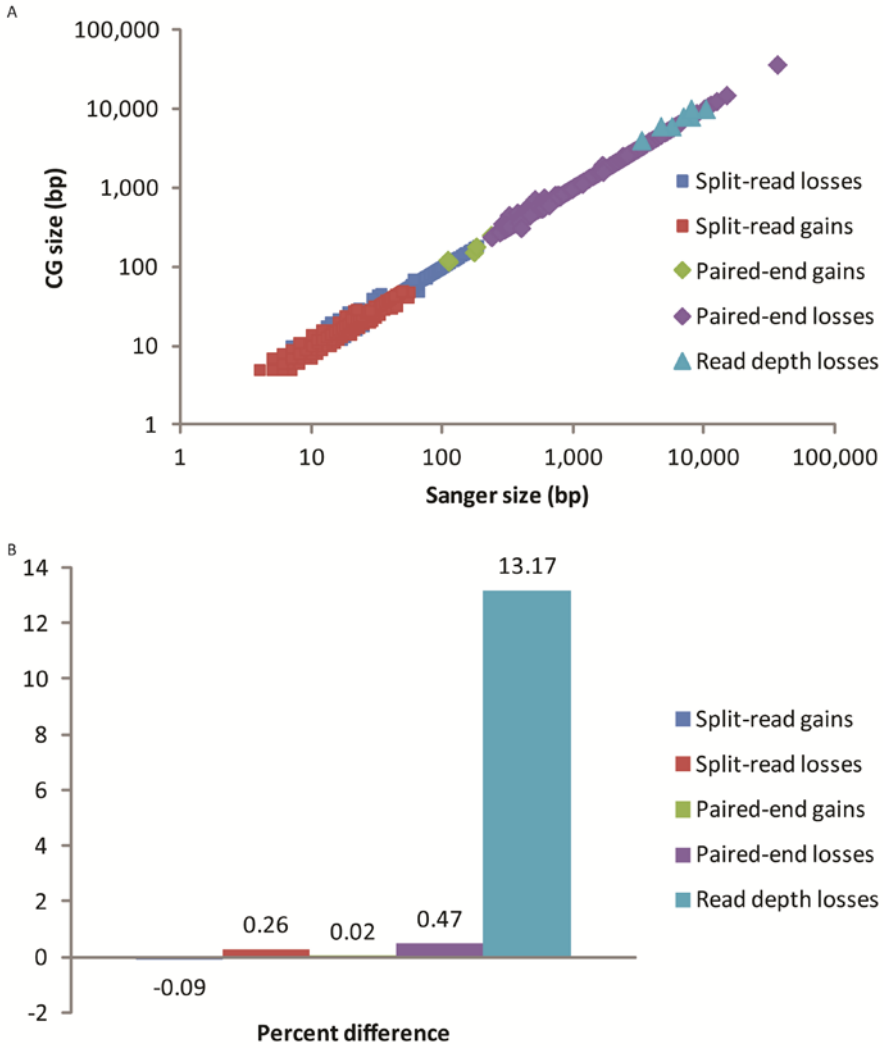


Figure V. 7. HuRef CG variant size estimation.
(A) shows the tight size correlation between HuRef CG variants (> 5 bp) and the corresponding breakpoint-refined Standard variants, and (B) displays the average percent size difference between HuRef CG and HuRef Standard calls.

Through an assessment of indels and CNVs discovered in the HuRef genome, one would notice that short read sequencing are still missing a notable number of variants, especially gains readily detected by the paired-end approach (Figure V. 3). To address this, I recommend generating libraries of multiple insert lengths. Even without changing the overall coverage, having multiple insert sizes should improved sensitivity: small libraries are better at calling small and localizing breakpoints; large insert libraries at calling large variants (Medvedev, et al., 2009). The deficiency in detecting variation in repeats is with short read length (Figure V. 5). With long reads, even ultra-long trinucleotide expansion can be effectively captured (Loomis, et al., 2013). Computationally, one should continue to apply multiple complementary strategies: split-read, paired-end, read depth, and one-end-anchor (Hajirasouliha, et al., 2010) (further discuss below). Future studies can also consider incorporating whole genome assembly comparison approach, as it can yield the greatest number, type and size range of variants (Table V. 2). However, current *de novo* assembly of short sequences is hampered by repeats. A possible solution is a hybrid assembly constructed by a mixture of shallow coverage (~5x) of mate-pair long-read sequencing with deeper coverage (~25x) of short-read sequencing (Schatz, et al., 2010). Alternatively, sequencing can be performed in conjunction with microarray (Pinto, et al., 2011) or optical mapping (Teague, et al., 2010) to detect large variation. In the latter case, besides determining the genomic position of long DNA fragments by optical map, one can sequence each isolated fragment, and map the reads to the corresponding position. This and other (e.g. Long Fragment Read (Peters, et al., 2012)) processes of complexity reduction should improve alignment and variation-discovery accuracy. Finally, some common variants (minor allele frequency >5%) that are missed by NGS could be imputed by nearby tag SNPs, but some rare variants would not be tagged; for example, approximately 20% of biallelic structural variants cannot be readily captured (Mills, et al., 2011b). Ultimately, if NGS is to become a primary technology in clinical laboratories (Gargis, et al., 2012), it will benefit from improvement, particularly in capturing indels, CNVs, inversions and more complex rearrangements that are associated with diseases (Mills, et al., 2011a; Tang and Amon, 2013).

V.B.2 Improvement of the HuRef variation map

As shown in Chapter II and the previous section, both the HuRef Standard and HuRef CG profiles contained false positives and false negatives. While the HuRef Standard had the advantage of having long and accurate reads generated from long mate-pairs, the HuRef CG benefited from having deep coverage. I believe one can improve the HuRef Standard variation map by incorporating NGS data to validate existing calls and to identify novel variation.

In Chapter, I show that there are currently many human genomes that have been sequenced using different NGS platforms. Furthermore, there is a plethora of software suites designed to analyze NGS data. While some are platform-specific or task-specific, others are platform-agnostic and multi-purpose. Tables V. 4 to V. 7 show some alignment, substitution and structural variation detection programs. In the future, one can apply some of the listed algorithms on HuRef NGS data to uncover additional variants. I anticipate that NGS will discover many new heterozygous variants currently missed due to shallow coverage. Hence, the entire HuRef variation size distribution curve will likely elevate.

Table V. 4. Alignment tools for NGS

Program*	Platform	Website
BFAST	Illumina/Life	http://sourceforge.net/apps/mediawiki/bfast/index.php?title=Main_Page
Bowtie	Illumina/Roche/Life	http://bowtie-bio.sourceforge.net
BWA	Illumina/Life	http://bio-bwa.sourceforge.net/bwa.shtml
CoronaLite	Life	http://solidsoftwaretools.com/gf/project/corona/
CABOG	Roche/Life	http://wgs-assembler.sf.net
ELAND/ELAND2	Illumina/Life	http://www.illumina.com/
EULER	Illumina	http://euler-assembler.ucsd.edu/portal/
Exonerate	Roche	http://www.ebi.ac.uk/~guy/exonerate
EMBF	Illumina	http://www.biomedcentral.com/1471-2105/10?issue=S1
GenomeMapper	Illumina	http://1001genomes.org/downloads/genomemapper.html
GMAP	Illumina	http://www.gene.com/share/gmap
Gnumap	Illumina	http://dna.cs.byu.edu/gnumap/
ICON	Illumina	http://icorn.sourceforge.net/
Karma	Illumina/Life	http://www.sph.umich.edu/csg/pha/karma/
LAST	Illumina	http://last.cbrc.jp/
LOCAS	Illumina	http://www-ab.informatik.uni-tuebingen.de/software/locas
Mapreads	Life	http://solidsoftwaretools.com/gf/project/mapreads/
MAQ	Illumina/Life	http://maq.sourceforge.net
MOM	Illumina	http://mom.csbc.vcu.edu/
Mosaik	Illumina/Roche/Life	http://bioinformatics.bc.edu/marthlab/Mosaik
mrFAST/mrsFAST	Illumina	http://mrfast.sourceforge.net/
MUMer	Life	http://mummer.sourceforge.net/
Nexalign	Illumina	http://genome.gsc.riken.jp/osc/english/dataresource/
Novocraft	Illumina	http://www.novocraft.com/
PerM	Illumina/Life	http://code.google.com/p/perm/
RazerS	Illumina/Life	http://www.seqan.de/projects/razers.html
RMAP	Illumina	http://rulai.cshl.edu/rmap
Segemehl	Illumina/Roche	http://www.bioinf.uni-leipzig.de/Software/segemehl/
SeqCons	Roche	http://www.seqan.de/projects/seqcons.html
SeqMap	Illumina	http://biogibbs.stanford.edu/*jiangh/SeqMap/
SHRiMP	Illumina/Roche/Life	http://compbio.cs.toronto.edu/shrimp
Slider/SliderII	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/slider
SOCS	Life	http://solidsoftwaretools.com/gf/project/socs/
SOAP/SOAP2	Illumina/Life	http://soap.genomics.org.cn
SSAHA/SSAHA2	Illumina/Roche	http://www.sanger.ac.uk/Software/analysis/SSAHA2
Stampy	Illumina	http://www.well.ox.ac.uk/~marting/
SXOligoSearch	Illumina	http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
SHORE	Illumina	http://1001genomes.org/downloads/shore.html
Vmatch	Illumina	http://www.vmatch.de/

* This table is adopted from (Zhang, et al., 2011).

Table V. 5. Single nucleotide variation detection programs designed for NGS.

Program*	Platform	Website
Atlas-SNP2	Roche/Illumina	http://www.hgsc.bcm.tmc.edu/cascade-tech-software-ti.hgsc
BOAT	Illumina	http://boat.cbi.pku.edu.cn/
DNA Baser	Roche	http://www.dnabaser.com/help/manual.html
DNAA	Roche/Illumina/ABI	http://sourceforge.net/projects/dnaa/
Galign	Illumina	http://shahamlab.rockefeller.edu/galign/galign.htm
GigaBayes/PbShort	Roche/Illumina	http://bioinformatics.bc.edu/marthlab/GigaBayes
GSNAP	Roche/Illumina	http://share.gene.com/gmap.
inGAP	Roche/Illumina	http://sites.google.com/site/nextgengenomics/ingap
ngs_backbone	Roche/Illumina	http://bioinf.comav.upv.es/ngs_backbone/index.html
Omixon Variant	ABI	http://www.omixon.com/omixon/index.html
PyroBayes	Roche	http://bioinformatics.bc.edu/marthlab/PyroBayes
Slider	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/slider
SNP-o-matic	Illumina	http://snpomatic.sourceforge.net
SNPSeeker	Illumina	http://www.genetics.wustl.edu/rmlab/
SNVMix	Illumina	http://compbio.bccrc.ca
SOAPsnp	Roche/Illumina/ABI	http://soap.genomics.org.cn
ssahaSNP	Illumina/Roche	http://www.sanger.ac.uk/Software/analysis/ssahaSNP
SVA	Illumina	http://www.svaproject.org/
SWA454	Roche	http://www.broadinstitute.org/science/programs/genome-biology/crd
VAAL	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd
VARiD	Roche/Illumina/ABI	http://compbio.cs.utoronto.ca/varid
VarScan	Roche/Illumina	http://genome.wustl.edu/tools/cancer-genomics

* This table is adopted from (Zhang, et al., 2011).

Table V. 6. Structural variation detection programs designed for NGS.

Program*	Platform	Website
BreakDancer	Roche/Illumina/Life	http://genome.wustl.edu/tools/cancer-genomics/
BreakDancer/BD- Mini	Roche/Illumina/Life	http://seqanswers.com/wiki/BreakDancer
Breakway	Roche/Illumina/Life	http://sourceforge.net/projects/breakway/files/
cnD	Illumina	http://www.sanger.ac.uk/resources/software/cnd.html
CNVer	Illumina	http://compbio.cs.toronto.edu/cnve
cnvHMM	Illumina	http://genome.wustl.edu/pub/software/cancer-genomics/cnvHMM/
CNVSeq	Roche	http://tiger.dbs.nus.edu.sg/CNV-seq/
GASV/GSV	Illumina	http://cs.brown.edu/people/braphael/software.html
Hydra	Illumina	http://code.google.com/p/hydra-sv/
MoDIL	Illumina	http://compbio.cs.toronto.edu/modil/
mrCaNaVaR	Roche/Illumina/Life	http://mrcanavar.sourceforge.net/
NovelSeq	Roche/Illumina/Life	http://compbio.cs.sfu.ca/strvar.htm
PEMer	Roche/Illumina/Life	http://sv.gersteinlab.org/pemer/
Pindel	Illumina	http://www.ebi.ac.uk/~kye/pindel/
PRISM	Illumina/Life	http://compbio.cs.toronto.edu/prism/
SegSeq	Illumina/Life	http://www.broadinstitute.org/
SOAPsv	Roche/Illumina/Life	http://soap.genomics.org.cn
Solid CNV tool	Life	http://solidsoftwaretools.com/gf/project/cnv/
Solid large Indel tool	Life	http://solidsoftwaretools.com/gf/project/large_indel/
SWT	Illumina	http://genome.wustl.edu/pub/software/cancer-genomics/GSTAT/
VariationHunter/VH-CR	Illumina	http://compbio.cs.sfu.ca/strvar.html
VARiD	Life	http://compbio.cs.utoronto.ca/varid

* This table is adopted from (Zhang, et al., 2011).

Table V. 7. Multi-task software suites designed for NGS.

Multi-task software packages	Website
BING	http://www.dinulab.org/bing
Bioscope	https://products.appliedbiosystems.com/ab/en/US/adirect/
CASAVA	http://www.illumina.com/software/
CGA Tools	http://www.completegenomics.com/analysis-tools/
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/
Geneious Pro	http://www.geneious.com/default,1246,NGS%20Assembly.sm
Geneus/GenoLogics	http://www.genologics.com/solutions/research-informatics/
Genomatix Genome Analyzer	http://www.genomatix.de/genome_analyzer.html
Genomic workbench/CLCbio	http://www.clcbio.com/index.php?id=1331
JMP Genomics	http://www.jmp.com/software/genomics/index.shtml
NextGENe/SoftGenetics	http://softgenetics.com/NextGENe.html
PacBio RS system	http://www.pacificbiosciences.com/products/software/
PaCGeE/PGI	http://personalgenomicsinstitute.org/index.php/
Partek GS/Partek	http://www.partek.com/partekgs
PASS	http://pass.cribi.unipd.it/cgi-bin/pass.pl?action=Download
Roche Analysis tools	http://454.com/products-solutions/analysis-tools/index.asp
RTG/Real Time Genomics	http://www.realtimegenomics.com/RTG-Software
SeqMan Ngen/DNASTAR	http://www.dnastar.com/t-products-seqman-ngen.aspx
TorrentSuite Software	http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Semiconductor-Sequencing/data_analysis/torrent_browser.html
VSRAP	http://sourceforge.net/apps/mediawiki/vancouvershortr/
Zoom	http://www.bioinformaticssolutions.com/products/zoom/index.php

* This table is adopted from (Zhang, et al., 2011).

V.B.3 Determine the genotype of CNVs

Global assessment of the genotype or the absolute copy number of CNVs has been challenging for microarrays. SNP array assays are originally designed to discriminate SNP alleles rather than copy number measurements. And by measuring the relative intensity ratios in CGH arrays, it is difficult to discern copy number of multi-copy duplications. Nevertheless, the ability to accurately predict copy number can enable making genotype and phenotype correlation. For example, an individual with higher copy number of the *CCL3L1* gene than the average of his/her ethnic background tends to have greater resistance to HIV infection (Gonzalez, et al., 2005). Alkan and colleagues used read depth information to predict the absolute copy number of segmental duplications and CNVs in two deeply sequenced genomes (Alkan, et al., 2009). They demonstrated that this approach can distinguish multi-copy number difference (e.g. copy number 5 versus 12), a feat which is not attainable by microarrays due to the saturation of fluorescence intensities. Importantly, genes with highly variable copy number change tend to reside in duplicated loci, thus highlighting the dynamic nature of these regions. Some of these genes correspond to rapidly evolving gene families such as the zinc finger and the Morpheus families. Moreover, advance in digital PCR technology is capable of determining the exact copy number count of DNA segment (Sykes, et al., 1992). While multiplexing is currently under development, the technology can be used for validating estimations generated by sequence count. Hence, CNV-genotyping should be a routine task in future studies.

V.B.4 Breakpoint refinement

As discussed in Chapter III, current array or sequence based studies can reliably detect gains and losses of DNA, but nonetheless their precise breakpoint information may not be readily available. Particularly, complex loci with repeats or segmental duplications are difficult to align to, and can cause spurious alignments. Therefore, signatures of reads that capture true breakpoints can be obscured by surrounding noisy alignments. Certainly longer read length can improve the accuracy of alignment, variant detection, and subsequent breakpoint refinement by local assembly (Li, et al., 2010b; Simpson, et al., 2009; Zerbino and Birney,

2008). In addition, more targeted approaches, or other creative high-throughput sequence capture methods such as sequence capture using probes/baits (Conrad, et al., 2010a) designed from variant junction libraries (Lam, et al., 2010) are needed to elucidate the underlying variant structures.

V.B.5 Detection and annotation of novel DNA sequences

There are many insertion events in the HuRef genome that are absent in the public reference assembly. Similar observations have also been reported in other studies (Hajirasouliha, et al., 2010; Kidd, et al., 2010b; Li, et al., 2010a; Wheeler, et al., 2008). It has been estimated that 19 to 40 Mb of sequences is missing in the reference. These sequences can represent insertions in the sequenced genome, or they can correspond to reference assembly gaps (Bovee, et al., 2008). These DNA fragments may have functional units such as enhancers, coding and other non-coding sequences. They may be polymorphic, exhibit population differentiation or individual-specific, and contribute to the phenotypic diversity and different disease susceptibility. Yet, since these sequences are absent in commercially available microarrays, and are typically not sought for in variation studies, our understanding of these sequences is noticeably less than other euchromatic sequences readily reported in existing genome browsers.

In Chapter II, a custom Agilent 244k CGH array was designed to search for evidence of copy number change in sequences present in the Celera assembly and not in the public reference assembly, and I demonstrated that these sequences were indeed polymorphic among a cohort of seven individuals.

One can also use computational method to detect novel insertion sequences using mate pair sequence data. In a genomic region upstream or downstream of site of a large insertion event, there should be an abundant number of mate-pair inserts where only one of the mates would align to the reference genome (Figure V. 8). Hence, to search for large insertion sites, one can look for loci where there is a significant number of these “one-end anchored” (OEA) inserts (the green reads in Figure V. 8).

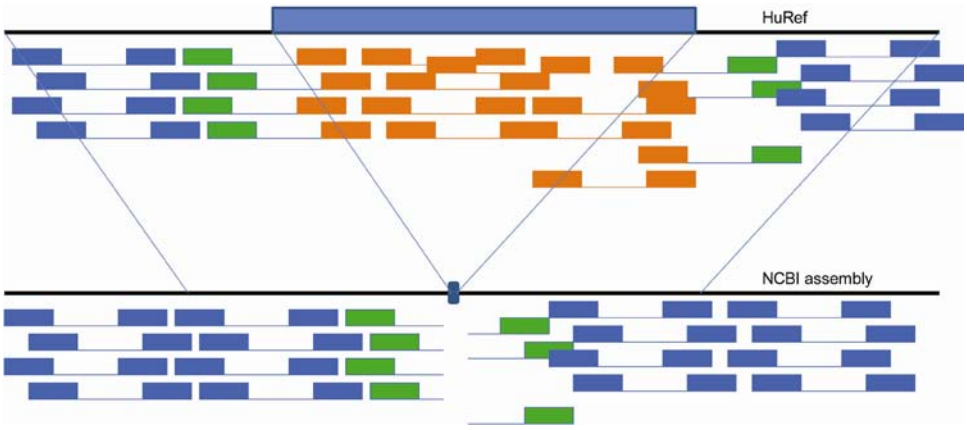


Figure V. 8. Schematic of detection of insertion by OEA mapping.
The presence of a sequence (thick blue box at top) in a sample, in this case HuRef, not present in the NCBI reference assembly would create a significant number of OEA inserts around an insertion breakpoint. The mapped end of the OEA read is colored in green. All unmapped reads are colored in orange while all other paired reads are colored in blue. This OEA signature can be used to identify insertion sequences in the reference genome.

Finally, whole genome *de* novo assembly can also detect and annotate these novel sequences; however, this is still challenging for current short read NGS technology. Future studies can use a combination of custom microarrays (Kidd, et al., 2010b; Pang, et al., 2010), OEA mapping approaches (Hajirasouliha, et al., 2010; Kidd, et al., 2010b), and local assembly to identify these DNA sequences, reveal their functional and structural importance, and close the remaining 271 gaps in the public assembly. Of course, these new sequences should then be incorporated into the reference. Because the reference assembly should ultimately encompass the longest chromosomal sequences, incorporating novel DNA from multiple studies, in order to represent all possible DNA in the human species (Feuk, et al., 2006a; Scherer, et al., 2007).

V.B.6 Inversion detection

In Chapter III, I show that the majority of large size (> 1 kb) inversions detected in the HuRef genome are flanked by homologous segmental duplications and interspersed repeats. These repeats can obscure mate-pair alignments. I envision that this problem of misalignment will be alleviated with improvement in NGS chemistry in generating longer reads from larger insert libraries. On the other hand, the detection of small inversions, which are less often flanked by homologous DNA, can be enhanced by having deeper coverage, thus increasing the number of DNA fragments covering variant breakpoints. Finally, understudied ultramicro-inversions may be captured by algorithms that search for strand-flipping alignments (Hara and Imanishi, 2011; Ye, et al., 2009).

V.C Structural variation *de novo* rate

The rate of formation has been known to differ among variation types. For example, the rate differs between SNPs and CNVs, as they are formed by different mutagenesis process (Table V. 4). Similarly, from Chapter III, I show that multiple mechanisms operate within even the broad categories of structural variation. So, I hypothesize that the *de novo* rate of formation is different for each mechanism. For instance, structural variants formed by replication processes such as FoSTeS or MMBIR would likely differ from those formed by recombination-based NAHR events. Replication errors tend to correlate with paternal age, but recombination ones do not (Zhang, et al., 2009a). Furthermore, NAHR depends on the

structure of other genomic architectures. In the case of 17q21.31 microdeletion syndrome, parents of patients with the 424 kb deletion carry a 900 kb inversion at the deleted locus (Koolen, et al., 2008). The deletion and inversion are flanked by segmental duplications, and the inversion contains the specific segmental duplication structure necessary to mediate the formation of the pathogenic deletion by NAHR during meiosis (Itsara, et al., 2010). So in this case, the rate of deletion-formation would vary between chromosomes that have the inversion and those that do not. In addition, the inversion is present in ~ 20 % of the European population but is rarer in other populations (Stefansson, et al., 2005). So recombination rate may also vary in different populations. Therefore, due to heterogeneity in local sequence structures and haplotype frequencies, our current estimate of $1.5x10^{-2}$ new CNVs per generation is likely an average of all mechanism types. Future *de* novo rate estimation should sub-divide by mechanism type, and consider the ethnicity of the samples.

Table V. 8. De novo mutation rate of various types of variation.

Type	Mutation rate (per genome per generation)	Reference	Size of variants studied	# of corresponding variants HuRef
SNV	70	(Conrad, et al., 2011)	1 bp	3,213,401
Small indel[*†]	3	(Lynch, 2010)	1 – 50 bp	581,280
Retrotransposition[**]	4.6×10^{-2}	(Stewart, et al., 2011)	30 – 6,250 bp	1,542
CNV	1.2×10^{-2}	(Conrad, et al., 2010b; Itsara, et al., 2010)	> 500 bp	4,072

[*] The rate excludes micro- and minisatellite loci. The study only examined 2,585 deletions and 903 insertions residing in 21 loci associated with autosomal dominant and 13 loci associated with X-linked disorders. The study also ignored indels whose length is divisible by three, and its reason was that those variants would leave codon reading frame intact and would have minimal phenotypic effects. The number of HuRef indel variants indicated in the last column excludes those characterized as microsatellite or minisatellite in Chapter III.

[†] Expansion and contraction of microsatellites has been independently examined at 2,477 autosomal loci, and the mutation rate is estimated to be between 2.7×10^{-4} to 10.0×10^{-4} per locus per generation (Sun, et al., 2012).

[**] This rate was calculated based on a map of 7,380 Alu, L1 and SVA detected in 185 samples. This study was based mainly on low-pass short-read sequencing data, thus explaining the relatively low number of retrotransposons detected. The coverage per sample was about 3.0 x. The number of HuRef retrotransposition has been determined in Chapter III.

V.D Towards a complete variation map of the human genome

A reader may notice that I have a recurring theme in this thesis, and that is the importance of having the complete catalog of variation. The HuRef data set, being the most complete to date, offers many unique opportunities: examine the strengths of different discovery methods, genotype the rarely studied insertions and inversions, quantify the relative proportion of mutational mechanisms for variants of different size, et cetera. None of these tasks would be possible, or at least the results might be less accurate, had I used less complete data with notable gaps. The importance to study all forms of variation is also recognized in other population (Altshuler, et al., 2010; Durbin, et al., 2010; Jakobsson, et al., 2008) and clinical studies (Berkel, et al., 2010; Sanders, et al., 2012). At the present moment, to get a complete set of structural variation, one cannot rely on SNP-based imputation, and has to employ multiple direct discovery approaches. So in the future, how will we be able to get a "complete" variation map of the human genome?

The coming third generation sequencing approach has the potential address some existing issues in variation-detection by NGS. There are two main characteristics to the third generation technology: PCR is not needed before sequencing; and sequencing signal is captured in real time. No pre-sequencing amplification enables shortening of DNA preparation time and elimination any systematic bias in PCR amplifications. Sequencing signal in real time means that the signal is captured during enzymatic reaction of adding nucleotide. Uninterrupted, DNA polymerase can incorporate multiple bases per second; hence natural long length DNA can be produced. There are two notable third generation sequencing methods, and they are the Pacific Bioscience's Single-molecule real-time (SMRT) method and Nanopore DNA sequencing. The average read length of PacBio RS machine is about 1.3 kb, while Nanopore can potentially reach over 5 kb read length (Liu, et al., 2012). Both lengths are significantly longer than what can be achieved by Sanger sequencing and NGS. Long read length enables placement of sequenced reads to their proper location, and that can subsequently improve variation discovery. It will be exciting to see if the third generation sequencing can detect all types of variation, thus potentially avoiding the need to use multiple approaches to find the full spectrum of variation.

How will one get a "complete" human variation map? I think that there are two prerequisites to generate such map, and they are 1) the availability of accurate sequencing and 2) the availability of genome sequence from a large number of individuals. To achieve the first prerequisite, the Archon Genomics X prize is set up to challenge the scientific community to radically improve sequencing technology. The participating teams will have to rapidly, accurately and economically sequence 100 human genomes (Kedes and Campany, 2011; Kedes, et al., 2011). Specifically, a $10 million prize will be awarded to the team to sequence the samples within 30 days with an error rate of 1 error per megabase, with 98 % genome coverage, identification of genetic variation, completely phased the variants, and at a cost of $1,000 per genome. The competition took place in January 3rd, 2013. Moreover, the 100 samples have been derived from genomes of centenarian samples, and the findings of the competition can potentially enhance our understanding to longevity and health.

Second, initiatives such as the 1000 Genomes Project and the Personal Genome Project (PGP) will enable the collection of variation information from many samples from the general public. The 1000 Genomes Project (www.1000genomes.org), whose goal is to sequence 2,500 genomes from 27 populations. The Personal Genome Project aims to enroll 100,000 volunteers from the general public (Ball, et al., 2012). In addition, the Personal Genome Project records very detailed phenotype information such as personal medical history, and that would enable the development of tools to correlate genomic information to phenotypes. The 1000 Genomes Project, Personal Genome Project as well as others will facilitate the continual accumulation of variation data, and in the near future we may find out the full extent genetic variation in the human DNA. This comprehensive catalogue of human genetic variants can in turn be used a reference for disease association.

V.E Personal genomics and medical relevance

Ultimately, what can we learn from sequencing healthy individuals, with no disease phenotype? We may be able to discover carrier status for incompletely penetrant dominant variants and recessive variants for monogenic disorders. We currently have limited idea on the impact on phenotype for the majority of variants, both substitution and structural variants. Nevertheless, those known variants associated with physical traits can be used for risk calculation for developing a disease, and many of these variants are genotyped by DTC

companies (Ng, et al., 2009). Yet, the odds ratio of most of these variants is low, thus providing limited predictive values. Better prediction can be done by incorporating genomic data with data such as diet, exercise and clinical characteristics. Currently, these analyses provide limited but useful information for an individual (Ashley, et al., 2010).

Since the publication of the first individual human genome (the HuRef genome) in 2007, there has been significant improvement in sequencing. Such advancement will surely continue at an even faster pace. However, the greatest challenge in the future is not in sequencing, but in the interpretation of the data. We still know little on the effects of most variation, as well as the genetic cause of many complex traits. First, we need a better understanding of the genome, in addition to protein-coding regions. Novel techniques now enable us to examine the regulatory landscape and three-dimensional DNA organization (ENCODE Project Consortium, et al., 2012). These could enhance our understanding of distal effects mediated by variants. Second, it is important to collect detailed human phenotypes, according to agreed upon standards, together with the deluge of genomic information. This availability of both data sets from a large number of individuals is fundamental to predict outcomes from sequences. Genetic information has the potential to improve the ability to direct lifestyle change and therapeutic selection. An example of this is can be seen with the HuRef individual. From his family history and from his genotype, Dr. Venter knows that he is at risk of cardiac problem, so he is proactively exercising and taking the cholesterol-lowering drug statin to address this condition (Venter, 2007). Yet despite such great expectations, we should remember that the effect of the vast majority of genetic variants is unlikely to be deterministic, as additional genetic, epigenetic and environmental interactions can influence an individual's phenotype.

The work in this thesis highlights numerous structural variation characteristics. It emphasizes the need to study the full complement of variation in personal genome, population and disease studies. The collective information gathered from analyzing the HuRef genome, some of which are reported in this thesis, can provide a good standard in the rapidly growing field. It contributes towards a greater understanding of the human genome, and ultimately will help unravel the association between genotypes and phenotypes.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56-65.

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: An approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res.

Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim D, Kim BC, Kim SY, Kim WY, Kim C and others. 2009. The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. Genome Res.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12(5):363-76.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O and others. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41(10):1061-7.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-10.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB and others. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52-8.

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. Hum Mol Genet 18(14):2555-66.

Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA and others. 2010. Clinical assessment incorporating a personal genome. Lancet 375(9725):1525-35.

Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, Angrist M, Bhak J, Bobe J, Callow MJ and others. 2012. A public resource facilitating clinical use of genomes. Proc Natl Acad Sci U S A 109(30):11920-7.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16(1):37-48.

Bansal V, Bafna V. 2008. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24(16):i153-9.

Baptista J, Mercer C, Prigmore E, Gribble SM, Carter NP, Maloney V, Thomas NS, Jacobs PA, Crolla JA. 2008. Breakpoint mapping and array CGH in translocations: comparison of a phenotypically normal and an abnormal cohort. Am J Hum Genet 82(4):927-36.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science 304(5675):1321-5.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573-80.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR and others. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53-9.

Berkel S, Marshall CR, Weiss B, Howe J, Roeth R, Moog U, Endris V, Roberts W, Szatmari P, Pinto D and others. 2010. Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. Nat Genet 42(6):489-91.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40(6):695-701.

Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tonnesen T, Carlberg BM, Pettersson U. 1995. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. Hum Mol Genet 4(4):615-21.

Bovee D, Zhou Y, Haugen E, Wu Z, Hayden HS, Gillett W, Tuzun E, Cooper GM, Sampas N, Phelps K and others. 2008. Closing gaps in the human genome with fosmid resources generated from multiple individuals. Nat Genet 40(1):96-101.

Buchanan JA, Scherer SW. 2008. Contemplating effects of genomic structural variation. Genet Med 10(9):639-47.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A and others. 2002. A human genome diversity cell line panel. Science 296(5566):261-2.

Carvalho CM, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, Shaw C, Peacock S, Pursley A, Tavyev YJ and others. 2009. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. Hum Mol Genet 18(12):2188-203.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP and others. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6(9):677-81.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE and others. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148(6):1293-307.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR and others. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat Genet 44(4):390-397.

Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 6(1):99-103.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S and others. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 106(45):19096-101.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR and others. 2011. Modernizing reference genome assemblies. PLoS Biol 9(7):e1001091.

Cohen SN, Chang AC, Boyer HW, Helling RB. 1973. Construction of biologically functional bacterial plasmids in vitro. Proc Natl Acad Sci U S A 70(11):3240-4.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. 2007. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35(6):2013-25.

Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, Turner DJ, Hurles ME. 2010a. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet 42(5):385-91.
Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV and others. 2011. Variation in genome-wide mutation rates within and between human families. Nat Genet 43(7):712-4.
Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P and others. 2010b. Origins and functional impact of copy number variation in the human genome. Nature 464(7289):704-12.
Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR and others. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1:131.
De S, Michor F. 2011. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat Biotechnol 29(12):1103-8.
Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G and others. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327(5961):78-81.
Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-73.
Edwards JH, Harnden DG, Cameron AH, Crosse VM, Wolff OH. 1960. A new trisomic syndrome. Lancet 1(7128):787-90.
ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J and others. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.
Fan HC, Wang J, Potanina A, Quake SR. 2011. Whole-genome molecular haplotyping of single cells. Nat Biotechnol 29(1):51-7.
Feuk L, Carson AR, Scherer SW. 2006a. Structural variation in the human genome. Nat Rev Genet 7(2):85-97.
Feuk L, Kalervo A, Lipsanen-Nyman M, Skaug J, Nakabayashi K, Finucane B, Hartung D, Innes M, Kerem B, Nowaczyk MJ and others. 2006b. Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. Am J Hum Genet 79(5):965-72.
Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet 1(4):e56.
Ford CE, Jones KW, Polani PE, De Almeida JC, Briggs JH. 1959. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). Lancet 1(7075):711-3.
Fox JL. 2008. What price personal genome exploration? Nat Biotechnol 26(10):1105-8.
Fudenberg G, Getz G, Meyerson M, Mirny LA. 2011. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat Biotechnol 29(12):1109-13.
Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M and others. 2010. Whole-genome sequencing and

comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet 42(11):931-6.
Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA and others. 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 30(11):1033-6.
Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL and others. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science 329(5993):841-5.
Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM and others. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329(5987):52-6.
Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, Ragusa A, Guerneri S, Selicorni A, Stumm M and others. 2002. Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. Am J Hum Genet 71(2):276-85.
Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ and others. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307(5714):1434-40.
Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108(29):11983-8.
Green PM, Bagnall RD, Waseem NH, Giannelli F. 2008. Haemophilia A mutations in the UK: results of screening one-third of the population. Br J Haematol 143(1):115-28.
Gupta R, Ratan A, Rajesh C, Chen R, Kim HL, Burhans R, Miller W, Santhosh S, Davuluri RV, Butte A and others. 2012. Sequencing and analysis of a South Asian-Indian personal genome. BMC Genomics 13(1):440.
Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC. 2010. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26(10):1277-83.
Hara Y, Imanishi T. 2011. Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. BMC Evol Biol 11:308.
Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S and others. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10(3):R32.
Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 5(1):e1000327.
Higgins AW, Alkuraya FS, Bosco AF, Brown KK, Bruns GA, Donovan DJ, Eisenman R, Fan Y, Farra CG, Ferguson HL and others. 2008. Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. Am J Hum Genet 82(3):712-22.
Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 19(7):1270-8.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23(2):254-67.
Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36(9):949-51.
Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, Walenz B, Shatkay H, Dew I, Miller JR and others. 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci U S A 101(7):1916-21.
Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. 2010. De novo rates and selection of large copy number variation. Genome Res 20(11):1469-81.
Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R and others. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181):998-1003.
Jiang Y, Wang Y, Brudno M. 2012. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. Bioinformatics 28(20):2576-83.
Johansson AC, Feuk L. 2011. Characterization of copy number-stable regions in the human genome. Hum Mutat 32(8):947-55.
Ju YS, Hong D, Kim S, Park SS, Lee S, Park H, Kim JI, Seo JS. 2010. Reference-unbiased copy number variant analysis using CGH microarrays. Nucleic Acids Res 38(20):e190.
Kedes L, Campany G. 2011. The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition. Nat Genet 43(11):1055-8.
Kedes L, Liu E, Jongeneel CV, Sutton G. 2011. Judging the Archon Genomics X PRIZE for whole human genome sequencing. Nat Genet 43(3):175.
Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336(6082):740-3.
Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M and others. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun 3:698.
Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12(4):656-64.
Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245(4922):1073-80.
Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, Rafiq MA, Qian C, Shago M, Pantano L and others. 2006. Genome assembly comparison identifies structural variants in the human genome. Nat Genet 38(12):1413-8.
Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F and others. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453(7191):56-64.
Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010a. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell 143(5):837-47.
Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G and others. 2010b. Characterization of missing human

genome sequences and copy-number polymorphic insertions. Nat Methods 7(5):365-71.
Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ and others. 2009. A highly annotated whole-genome sequence of a Korean individual. Nature 460(7258):1011-5.
Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE and others. 2011. Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29(1):59-63.
Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M and others. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum Mol Genet 20(10):1916-24.
Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM. 1987. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell 50(3):509-17.
Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. Semin Cancer Biol 20(4):211-21.
Koolen DA, Sharp AJ, Hurst JA, Firth HV, Knight SJ, Goldenberg A, Saugier-Veber P, Pfundt R, Vissers LE, Destree A and others. 2008. Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. J Med Genet 45(11):710-20.
Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L and others. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420-6.
Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet 91(6):1033-40.
Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K and others. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40(10):1253-60.
Lam HY, Clark MJ, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ and others. 2012. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30(1):78-82.
Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol 28(1):47-55.
Lam KW, Jeffreys AJ. 2006. Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. Proc Natl Acad Sci U S A 103(24):8921-7.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W and others. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921.
Lee C, Scherer SW. 2010. The clinical context of copy number variation in the human genome. Expert Rev Mol Med 12:e8.
Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131(7):1235-47.

Lee S, Cheran E, Brudno M. 2008. A robust framework for detecting structural variations in a genome. Bioinformatics 24(13):i59-67.
Lejeune J, Gautier M, Turpin R. 1959. [Study of somatic chromosomes from 9 mongoloid children]. C R Hebd Seances Acad Sci 248(11):1721-2.
Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G and others. 2007. The diploid genome sequence of an individual human. PLoS Biol 5(10):e254.
Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J and others. 2010a. Building the sequence map of the human pan-genome. Nat Biotechnol 28(1):57-63.
Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K and others. 2010b. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20(2):265-72.
Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364.
Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, Hagerman PJ. 2013. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome Res 23(1):121-8.
Lubs HA. 1969. A marker X chromosome. Am J Hum Genet 21(3):231-44.
Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA and others. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362(13):1181-91.
Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A 107(3):961-8.
MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N and others. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 72(6):971-83.
Maher B. 2008. Personal genomes: The case of the missing heritability. Nature 456(7218):18-21.
Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. Proc Natl Acad Sci U S A 74(2):560-4.
McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A and others. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40(10):1166-74.
McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC and others. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19(9):1527-41.
McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC. 2012. nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Res 22(11):2250-61.
Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6(11 Suppl):S13-20.
Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16(9):1182-90.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C and others. 2011a. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res 21(6):830-9.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK and others. 2011b. Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59-65.

Myers R. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9(4):e1001046.

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. 2008. Genetic variation in an individual human exome. PLoS Genet 4(8):e1000160.

Ng PC, Murray SS, Levy S, Venter JC. 2009. An agenda for personalized medicine. Nature 461(7265):724-6.

Nowell PC, Hungerford DA. 1961. Chromosome studies in human leukemia. II. Chronic granulocytic leukemia. J Natl Cancer Inst 27:1013-35.

Nussbaum RL, McInnes RR, Willard HF, Thompson MW, Hamosh A. 2007. Thompson & Thompson genetics in medicine. Philadelphia: Saunders/Elsevier.

Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui LC and others. 2001. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet 29(3):321-5.

Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L and others. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. Genome Res 21(1):33-46.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC and others. 2010. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11(5):R52.

Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP and others. 2010. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet.

Patau K, Smith DW, Therman E, Inhorn SL, Wagner HP. 1960. Multiple congenital anomaly caused by an extra autosome. Lancet 1(7128):790-3.

Patowary A, Purkanti R, Singh M, Chauhan RK, Bhartiya D, Dwivedi OP, Chauhan G, Bharadwaj D, Sivasubbu S, Scaria V. 2012. Systematic analysis and functional annotation of variations in the genome of an Indian individual. Hum Mutat 33(7):1133-40.

Pauling L, Itano HA, et al. 1949. Sickle cell anemia a molecular disease. Science 110(2865):543-8.

Pennisi E. 2007. Breakthrough of the year. Human genetic variation. Science 318(5858):1842-3.

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA and others. 2008. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82(3):685-95.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R and others. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256-60.
Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J and others. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487(7406):190-5.
Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R and others. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29(6):512-20.
Pinto D, Marshall C, Feuk L, Scherer SW. 2007. Copy-number variation in control population cohorts. Hum Mol Genet 16 Spec No. 2:R168-73.
Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S. 2008. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics 24(3):309-18.
Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. Nat Biotechnol 27(9):847-50.
Quinlan AR, Hall IM. 2012. Characterizing complex structural variation in germline and somatic genomes. Trends in Genetics 28(1):43-53.
Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T and others. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334(6052):94-8.
Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R and others. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463(7282):757-62.
Ray PN, Belfall B, Duff C, Logan C, Kean V, Thompson MW, Sylvester JE, Gorski JL, Schmickel RD, Worton RG. 1985. Cloning of the breakpoint of an X;21 translocation associated with Duchenne muscular dystrophy. Nature 318(6047):672-5.
Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W and others. 2006. Global variation in copy number in the human genome. Nature 444(7118):444-54.
Richard GF, Paques F. 2000. Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1(2):122-6.
Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL and others. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245(4922):1066-73.
Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M and others. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328(5978):636-9.
Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N and others. 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 245(4922):1059-65.
Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M and others. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348-52.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ and others. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-7.
Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL and others. 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485(7397):237-41.
Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463-7.
Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, Andraws N, Patterson ML, Krivohlavek LA, Fellis J and others. 2012. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med 4(154):154ra135.
Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. Genome Res 20(9):1165-73.
Scherer SW, Cheung J, MacDonald JR, Osborne LR, Nakabayashi K, Herbrick JA, Carson AR, Parker-Katiraee L, Skaug J, Khaja R and others. 2003. Human chromosome 7: DNA sequence and biology. Science 300(5620):767-72.
Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L. 2007. Challenges and standards in integrating surveys of structural variation. Nat Genet 39(7 Suppl):S7-15.
Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J and others. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463(7283):943-7.
Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M and others. 2004. Large-scale copy number polymorphism in the human genome. Science 305(5683):525-8.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117-23.
Small K, Iber J, Warren ST. 1997. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. Nat Genet 16(1):96-9.
Smit AF. 1996-2010. RepeatMasker Open-3.0.
Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG and others. 2005. A common inversion under selection in Europeans. Nat Genet 37(2):129-37.
Stepanov VA. 2010. Genomes, populations and diseases: ethnic genomics and personalized medicine. Acta Naturae 2(4):15-30.
Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA and others. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144(1):27-40.
Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP and others. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet 7(8):e1002236.
Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C and others. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813):848-53.

Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D and others. 2012. A direct characterization of human mutation based on microsatellites. Nat Genet.

Sykes PJ, Neoh SH, Brisco MJ, Hughes E, Condon J, Morley AA. 1992. Quantitation of targets for PCR by use of limiting dilution. Biotechniques 13(3):444-9.

Tang YC, Amon A. 2013. Gene copy-number alterations: a cost-benefit analysis. Cell 152(3):394-405.

Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM and others. 2010. High-resolution human genome structure by single-molecule analysis. Proc Natl Acad Sci U S A 107(24):10848-53.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437(7063):1299-320.

Tong P, Prendergast JG, Lohan AJ, Farrington SM, Cronin S, Friel N, Bradley DG, Hardiman O, Evans A, Wilson JF and others. 2010. Sequencing and analysis of an Irish human genome. Genome Biol 11(9):R91.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D and others. 2005. Fine-scale structural variation of the human genome. Nat Genet 37(7):727-32.

Venter JC. 2007. A life decoded : my genome ; my life. New York: Viking.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA and others. 2001. The sequence of the human genome. Science 291(5507):1304-51.

Wang J, Fan HC, Behr B, Quake SR. 2012. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell 150(2):402-12.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J and others. 2008. The diploid genome sequence of an Asian individual. Nature 456(7218):60-5.

Warburton D. 1991. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. Am J Hum Genet 49(5):995-1013.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT and others. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872-6.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA and others. 2009. Mobile elements create structural variation: Analysis of a complete human genome. Genome Research 19(9):1516-1526.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25(21):2865-71.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5):821-9.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009a. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451-81.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. 2009b. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41(7):849-53.
Zhang J, Chiodini R, Badr A, Zhang G. 2011. The impact of next-generation sequencing on genomics. J Genet Genomics 38(3):95-109.
Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW. 2006. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res 115(3-4):205-14.