Assessing the Functional Impact of Short Indels in Individual Human Genomes

David Goode[1], Dmitri Petrov[2] and Arend Sidow[1,3]

Departments of Genetics[1], Biological Sciences[2] and Pathology[3], Stanford University

Abstract

While SNVs and large structural variants in the human genome have been extensively studied, little is known about the functional impact of short insertions and deletions (indels) at the individual genome level. Using evolutionary constraint as a proxy for function, we assessed the potential functional impact of the short (<11 bp) indels ascertained from five individual genomes (a Chinese, a Korean, a Yoruba, a Khoisan and a Caucasian American). Our analysis reveals that there are at least 32,000 functional small indels in a single human genome. Despite differences between individuals in indel ascertainment, we observe consistent and strong selection against small indels in constrained sequences within each individual, demonstrating their functional importance. The strength of selection against short deletions scales with the number of constrained sites deleted, indicating that deletions that affect a greater number of constrained sites are more deleterious. Our results indicate that short indels potentially make a significant contribution to an individual's functional genetic variation, especially on non-coding sequences, and suggest that the functional importance of small indels deserves consideration during human genome sequencing studies.

Short Indel Data

Individual	Population	Sequencing method	# of Indels <=10 bp in length	% Insertions /Deletions
NA18507	Yoruban	Illumina	498,564	44.6%/55.4%
KB1	Khoisan	454	453,617	44.1%/55.9%
YH	Chinese	Illumina	135,174	45.1%/54.9%
SJK	Korean	Illumina	339,097	48.8%/51.2%
Venter	Caucasian American	Sanger	750,120	56.3%/46.4%

Table 1: Number of short (1-10 bp) indels detected in each individual genome. Variation in indel numbers is caused by sequencing read length and coverage depth, with longer read lengths and greater coverage allowing detection of more indels.

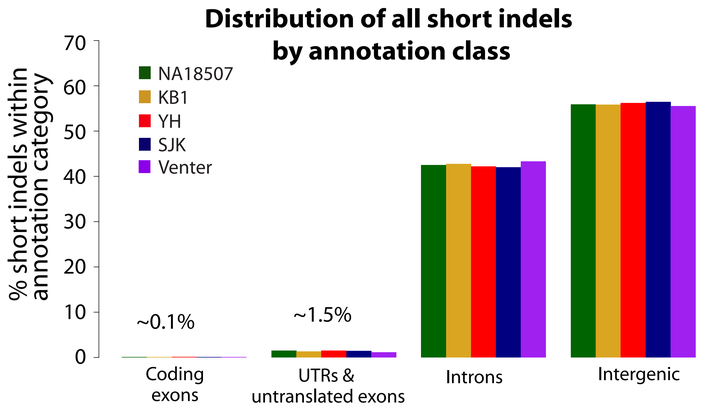

Figure 1: Percentage of short indels in each annotation class for all 5 individuals. Despite differences in indel ascertainment between individuals, short indels in each genome have the same distribution with respect to annotation. Thus, analyses performed using data from these 5 genomes should be directly comparable.

Evolutionary Constraint on Short Indels

Negative selection against indels in constrained elements

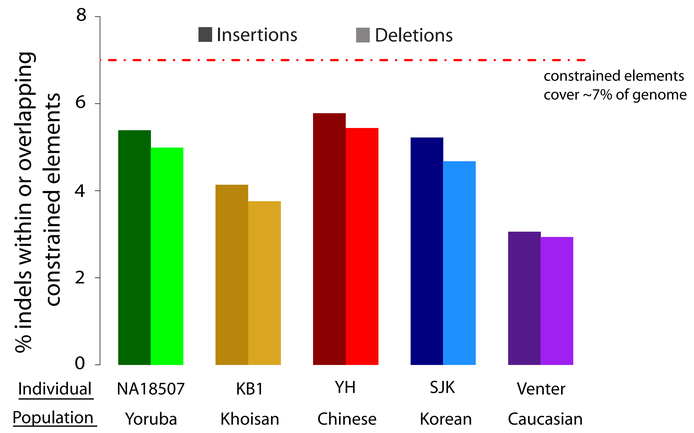

Figure 2: Percentage of insertions and deletions in each individual genome that overlap or are found within a constrained element. Both insertions (darker bars) and deletions (lighter bars) are depleted from constrained elements.

Deletion of constrained sites is strongly selected against, for both short and long indels

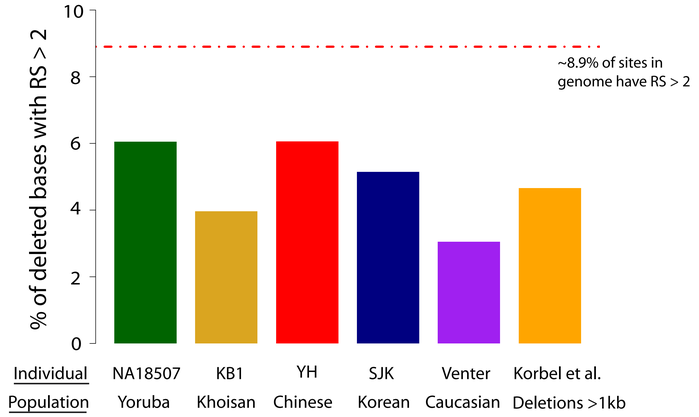

Figure 3: Percentage of all sites deleted by short (<=10 bp) deletions that are constrained (RS > 2) in each individual genome and in a set of verified human deletions >1kb (Korbel et al). The strength of selection against deleting sites that are under constraint is consistent for both short and long deletions.

Basis of Selection On Short Deletions

Strength of selection against deletions scales with the number of constrained sites deleted

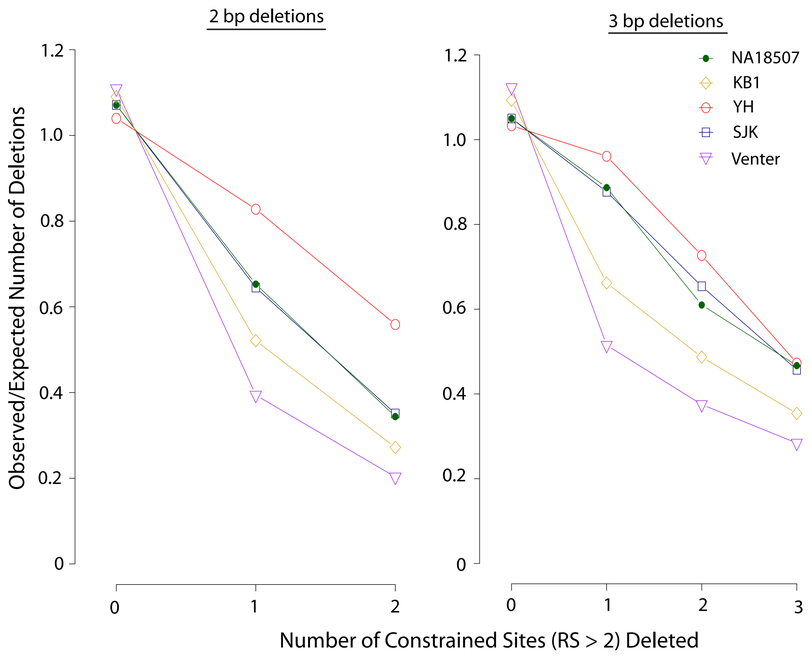

Figure 4: Ratio of the number of 2bp and 3 bp deletions that delete 0, 1, 2 or 3 constrained sites to the number expected from the genome-wide distribution of RS scores. This ratio declines as the number of constrained sites deleted increases, indicating that deletions which affect a greater number of constrained sites are more strongly selected against.

Potential Functional Impact Of Short Indels

Individual	# of insertions in CEs	# of deletions in CEs	# indels affecting CEs	# deletions affecting sites with RS > 2	# indels affecting constrained elements and/or constrained sites
YH	2,053	2,983	**5,036**	4,850	**6,903**
KB1	6,124	7,973	**14,097**	13,650	**19,774**
NA18507	9,577	12,745	**22,322**	23,327	**32,904**
SJK	5,044	8,055	**13,099**	15,220	**20,264**
Venter	8,858	8,805	**17,663**	17,276	**26,134**

Table 2: Number of potentially functional indels detected in each individual genome. Insertions are considered potentially functional if they occur within a constrained element, deletion if they delete any part of a constrained element or any constrained sites.

Distribution of putatively functional indels by annotation class

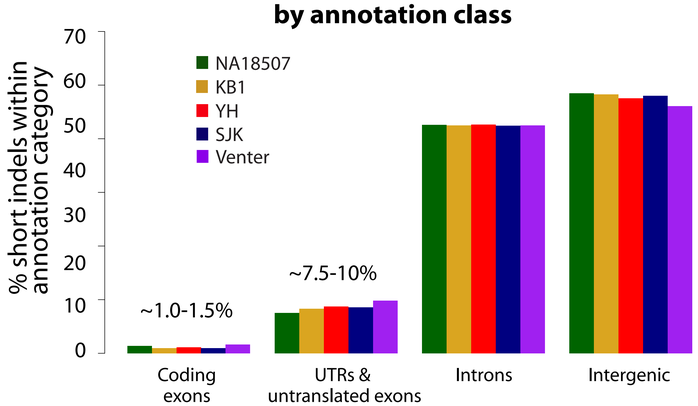

Figure 5: Percentage of putatively functional short indels in each annotation class for all 5 individuals. Non-coding sequences harbor the majority of putatively functional short indels, with a noticeable over-representation in the untranslated segments of transcripts.