DNA Repair 8 (2009) 579–584

Contents lists available at ScienceDirect

DNA Repair

journal homepage: www.elsevier.com/locate/dnarepair

Population-specific variation in haplotype composition and heterozygosity at the POLB locus

Jennifer Yamtich [b,a], William C. Speed [b], Eva Straka [b], Judith R. Kidd [b], Joann B. Sweasy [b,a], Kenneth K. Kidd [b,*]

[a] Departments of Therapeutic Radiology, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520, United States
[b] Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520, United States

ARTICLE INFO

Article history:
Received 7 July 2008
Received in revised form 9 December 2008
Accepted 9 December 2008
Available online 23 January 2009

Keywords:
POLB
Alleles
Haplotypes
Single nucleotide polymorphisms
Variation
Population genetics

ABSTRACT

DNA polymerase beta plays a central role in base excision repair (BER), which removes large numbers of endogenous DNA lesions from each cell on a daily basis. Little is currently known about germline polymorphisms within the POLB locus, making it difficult to study the association of variants at this locus with human diseases such as cancer. Yet, approximately thirty percent of human tumor types show variants of DNA polymerase beta. We have assessed the global frequency distributions of coding and common non-coding SNPs in and flanking the POLB gene for a total of 14 sites typed in approximately 2400 individuals from anthropologically defined human populations worldwide. We have found a marked difference between haplotype frequencies in African populations and in non-African populations.

1. Introduction

Cancer cells have a higher mutational load than can be accounted for by the normal cellular mutation rate, leading to the long-held hypothesis that there is a mutator phenotype in cancer (reviewed in [1]). Mutants with reduced fidelity of DNA repair and replication enzymes are important for the acquisition of this phenotype (reviewed in [2]). Genomic stability is challenged by the constant presence of endogenous and exogenous DNA damaging agents, some of which result in the production of 20,000 spontaneous apurinic/apyrimidinic (AP) sites per cell per day [3]. These cytotoxic lesions are the substrate for the base excision repair (BER) pathway, the DNA repair pathway primarily responsible for repairing lesions caused by reactive oxygen species and alkylating agents. DNA polymerase beta (Pol β) is a key enzyme in BER, acting as the main gap-filling enzyme and providing essential deoxyribose phosphate (dRP) lyase activity (reviewed in [4]).

Several tumor-specific single amino acid substitution mutants have been identified in pol β [5]. While it has been established that tumor-specific mutants may play a role in the etiology of cancer, little is currently known about the germline polymorphisms in pol β. dbSNP lists 189 single nucleotide polymorphisms (SNPs) in approximately 33 kb of the POLB locus, only 4 of which occur in the coding sequence (geneID:5423). Additionally, Mohrenweiser et al. [6] resequenced DNA repair genes in healthy individuals and identified three non-synonymous amino acid substitutions in pol β. To date, it is not known if any of these germ line SNPs overlap with the tumor-specific mutations. Only a few studies have tried to associate this germ-line variation in the POLB gene with altered cancer risk or survival, often with inconclusive results [7–11].

Using SNPs from both intronic and coding regions, previous studies have looked for association between variation at the POLB locus and cancer predisposition or survival. Several studies have found an association between POLB alleles and colorectal, breast, and bladder cancer risk [7–9] and lung and pancreatic cancer survival [10,11]. However, these studies are plagued by recurring limitations. First, the statistical power of the study is often limited by low frequencies of the variant alleles [7] and/or a lack of homozygotes in the study population [11], since one expects the effect of a variant to be greater in the homozygotes lacking a 'normal' allele. Second, many studies do not use haplotypes in their analysis, making it difficult to determine if the causative variant is the polymorphism being studied or another polymorphism in linkage disequilibrium with the genotyped marker. However, interactions of SNPs in the absence of strong LD could be missed since neither SNP alone would show as strong a signal as the haplotype. Thus haplotypes would be especially useful in studies such as in [8,9] which suggest that increased risk occurs with the presence

* Corresponding author. Tel.: +1 203 785 2654; fax: +1 203 785 6568.
E-mail address: kenneth.kidd@yale.edu (K.K. Kidd).

1568-7864/$ – see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.dnarep.2008.12.005

of multiple gene variants or as in [10] which suggests that being homozygous for the variant allele of either polymorphism alters survival.

One way to resolve the above limitations is through the determination of global population-specific allele frequencies for multiple DNA repair polymorphisms and the haplotypes comprised of the allelic combinations of the polymorphisms of each gene. Such data would be very informative. First, while variants in a single gene may be important, combinations of polymorphisms in different DNA repair genes could have unique phenotypes relevant to human disease, such as cancer predisposition. Global population-specific allele and haplotype frequencies will enable the determination of which allelic combinations, both in a specific gene and in multiple DNA repair genes, are most likely to occur. Second, this information would be useful in the design of biochemistry experiments aimed at testing the function of certain polymorphisms. By determining the odds of whether any two polymorphisms for the same gene or in the same pathway are likely to occur together, one can prioritize which combinations of variants to study. Finally, haplotype analysis will provide an evolutionary perspective on how variants have arisen in a gene. The ability to root the haplotype phylogeny, using the primate data, and the ability to infer a haplotype's relative age, based on distance from the root and global distribution, allows us to determine the likelihood of additional accumulated variation on a specific haplotype, given a constant mutation rate and absence of selective pressures. A more recent chromosome, e.g. a high-frequency recombinant haplotype seen only in North America, has had less time to accumulate additional variation than a chromosome close to the ancestral haplotype, found at moderate worldwide frequency.

As an initial study toward a full understanding of POLB variation globally, we have assessed the global frequency distributions of coding and common non-coding SNPs in and flanking the POLB gene for a total of 14 sites typed in approximately 2400 individuals from anthropologically defined human populations worldwide. These SNPs showed strong linkage disequilibrium. African populations had 3–4 common (>10%) haplotypes, while all non-African populations had a single common (generally >80%) haplotype and some populations had 1–2 others at >5%. The common non-African haplotype occurred at 20% or less in African populations. Three previously identified coding SNPs were not seen in any individuals worldwide, while two were generally low frequency and occurred in specific populations. Further resequencing of DNA repair genes in African individuals is warranted, as our results showed that several haplotypes common in African populations are rare or unobserved outside of Africa.

Table 1
Samples studied.

Geographic region	Population	N
African	Biaka Pygmy	68
	Mbuti Pygmies	39
	Yoruba	78
	Ibo	48
	Hausa	39
	Chagga	45
	Masai	21
	Sandawe	40
	African American	90
	Ethiopian Jews	32
SW Asian	Yemenite Jews	43
	Druze	102
	Samaritans	41
European	Ashkenazi Jews	81
	Adygei	54
	Chuvash	42
	Hungarians	89
	Russians, Archangel	33
	Russians, Vologda	48
	Finns	36
	Danes	51
	Irish	116
	European Americans	92
NW Asian	Komi Zyriane	47
	Khanty	50
S Central Asian	Keralite, S. India	30
NE Asian/Siberian	Yakut	51
from Pacific	Nasioi Melanesians	23
Islands	Micronesians	37
E Asian	Laotians	119
	Cambodians	25
	Chinese, San Francisco	50
	Chinese, Taiwan	50
	Hakka	41
	Koreans	54
	Japanese	51
	Ami	40
	Atayal	42
N American	Cheyenne	56
	Pima, Arizona	51
	Pima, Mexico	53
	Maya	52
S American	Quechua	23
	Ticuna	65
	Rondonian Surui	47
	Karitiana	57
Total		2442

For each geographic region, the population name and number of individuals in the sample are given. Descriptions of the populations and of the specific collections of samples can be found in ALFRED (http://alfred.med.yale.edu).

2. Methods

2.1. DNA sample collection and purification

The 2442 samples studied are listed in Table 1 by population name and grouped by geographic region. All samples were collected from apparently healthy adults; except in rare cases, no phenotype data were collected. These anonymous samples were collected with informed consent under protocols approved by the relevant committees at the various institutions and countries involved. DNA was extracted using standard phenol/chloroform purification from lymphoblastoid cell lines established and/or maintained by JRK at Yale.

2.2. SNP selection and analysis

SNPs were chosen based on several criteria: their worldwide frequency, their frequency distribution among populations, and their putative functional relevance (coding SNPs, 3′ UTR SNPs, etc.). All SNPs were genotyped by the TaqMan method (Applied Biosystems, www.appliedbiosystems.com) using commercial reagents and protocols. Four of the SNPs (Gln8Arg, Lys289Met, rs2953993, and rs3136804) required assays to be custom designed by Applied Biosystems (primer/probe sequences available on request). Ancestral state was determined by genotyping of 15 primates (3 gorilla, 3 chimpanzee, 3 orangutan, 3 bonobo, 3 gibbons) and comparison of human sequence to chimpanzee genomic reference sequence (genome.ucsc.edu).

2.3. PCR-RFLP analysis

Restriction fragment length polymorphism PCR (PCR-RFLP) was used to confirm the genotypes of the Lys289Met polymorphism of 89 samples of European American ancestry as in [9]. Briefly, each 20 μl PCR reaction contained 10 ng genomic DNA, 1.25 U Taq polymerase in 1 × PCR buffer, 1.5 mM $MgCl_2$, 50 μM dNTPs and 250 nM each primer. Forward primer sequence was