Editorial　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　**Open Access**

Pharmacogenomics of Simple Repeats: How Do You Solve a Problem like VNSR?

Evgeny Krynetskiy[*]

Temple University School of Pharmacy and Jayne Haines Center for Pharmacogenomics and Drug Safety, Philadelphia, USA

[*]**Corresponding author:** Evgeny Krynetskiy, Temple University School of Pharmacy and Jayne Haines Center for Pharmacogenomics and Drug Safety, Philadelphia, PA 19140, USA, E-mail: ekrynets@temple.edu

Editorial

The next big endeavor after the Human Genome Project is, of course, the Human Phenome Project, in other words, the ability to deduce the person's phenotype from the known genome, transcriptome, epigenome, and other …omic data. While certain phenotypes are easily predicted based on simple genotype-phenotype relationships such as Mendelian traits, many phenotypes including physical appearance, disease susceptibility, and physiological profile of an organism are still beyond our grasp. Luckily, Pharmacogenomics works with much simpler phenotypes. The phenotype in a PGx study is a physiological reaction to a chemical stimulus, usually a medication. Therefore, a pharmacogenomic model is the first step toward predicting macro scale characteristics of an organism based on its genome.

Toward this goal, the comparison of genomes between individuals, along with their medical records is performed. A little more than a decade after the first human genome was dissected, now the number of sequenced human genomes is counted in thousands, and the technological advances in the genome sequencing and analysis are amazing. Modern massive parallel analysis technologies, including Next Generation Sequencing (NGS), transcriptome, and proteome analysis, together with epigenomic and metabolomic studies, provide an unprecedented volume of information about the cellular events. Genetic variants including single nucleotide polymorphisms (SNPs), copy number variation (CNV), and indels, are collected, catalogued, and become easily accessible for analysis thus providing readily identifiable markers for a pharmacogenomic project.

But, there is another class of variable genetic elements that so far has not been adequately addressed in pharmacogenetic studies. While seemingly boring, the simple repeats are highly polymorphic, and are likely to play an important role in chromatin structure, dynamics, and gene regulation. Variable number of short repeats (VNSR) have been demonstrated to alter levels of gene expression, and therefore can have significant effect on phenotype. Analysis of VNSR in the human genome revealed a high density of repeat sequences around the transcription start sites (TSS) in multiple genes, and demonstrated that these microsatellites are statistically associated with promoters. Analysis of microsatellite distribution in the human genome identified $(A/T)_n$ and $(AC/GT)_n$ as two most common motifs within 5 kB of the TSS [1].

The modern genome analysis technique provides excellent tools to deal with the heterogeneous stretches of DNA, but is not as effective when the homogeneous or simple repeat regions of DNA are analyzed [2]. Though PCR-based estimation of VNSR length is straightforward, there are several reasons why bioinformatics analysis of simple repeats is not so simple, after all. The first problem is relatively short reads generated by NGS. To confidently map the simple dinucleotide repeat within the genome, the entire repeat region should be sequenced in one run, along with flanking regions long enough to provide exact mapping within the genome. With many repeats exceeding 150-200 bp, much of NGS whole genome data simply lack this information. The second problem is the propensity of the DNA polymerases to slip on simple repeats. This fact is demonstrated by PCR reactions where the amplification of dinucleotide repeats results in formation of heterogeneous products [3]. Therefore, amplification across a single VNSR locus produces multiple reads of variable lengths, an uncertainty which cannot be resolved by increasing number of reads through this region. Third, the assembler software demonstrate varying accuracy in aligning VNSR, and may require further optimization specifically for analysis of monotonous repeats [4]. Without a practical solution of assigning the exact length to the VNSR, the assembling algorithms should deal with these fuzzy data. Finally, this data should be mapped on the reference genome, without duplicating or compressing the repeats [2]. Without resolving these issues, significant amount of information about the length of short repeats in the genome may be lost within VNSR regions.

A number of studies evidence the effect of VNSR such as $(AC/GT)_n$ on phenotype, important examples including epidermal growth factor receptor *EGFR*, N-methyl-D-aspartate receptor subunit *GRIN2A*, heme oxygenase *HMOX1*, signal transducer and activator of transcription *STAT6*, UDP glucuronosyltransferase *UGT1A6*, and von Willebrand factor *VWF*. These genes contain $(AC/GT)_n$ repeats in promoter, exon, and intron regions. The number of human genes with long ($n\geq25$) $(AC/GT)_n$ repeats within 1 kB region upstream TSS exceeds 700, with many important, pharmacogenetically relevant genes such as ABC and SLC transporters, drug metabolizing enzymes, and drug targets. Multiple *in vitro*, *in vivo*, and clinical studies found a relationship between the lengths of VNSR and gene expression levels, though the underlying mechanisms remain obscure. In addition to SNP, CNV, and indels, the VNSR is yet another type of genetic variability directly related to gene expression level, and therefore can be a contributing factor to pharmacogenetic phenotype. This poorly characterized class of variable elements in the human genome clearly deserves attention from the pharmacogenetic community.

References

1. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One 8: e54710.

2. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, et al. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res 22: 557-567.

3. Imle P (2005) Fluorescent-based fragment size analysis. In: Innocenti F (ed) Pharmacogenomics: Methods and Protocols. Humana Press Inc., Totowa, NJ.

4. McIver LJ, Fondon JW 3rd, Skinner MA, Garner HR (2011) Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. Genomics 97: 193-199.