ПРИЛОЖЕНИЯ

к статье И.В. Чадаевой, Д.А. Рассказова, Е.Б. Шарыповой, И.А. Драчковой, Е.А. Ощепковой,
Л.К. Савинковой, П.М. Пономаренко, М.П. Пономаренко, Н.А. Колчанова, В.А. Козлова
«Кандидатные SNP-маркеры ревматоидного полиартрита, которые могут достоверно изменять
сродство ТАТА-связывающего белка к промоторам генов человека»

SUPPLEMENTARY MATERIALS

to the article I.V. Chadaeva, D.A. Rasskazov, E.B. Sharypova, I.A. Drachkova, E.A. Oshchepkova,
L.K. Savinkova, P.M. Ponomarenko, M.P. Ponomarenko, N.A. Kolchanov, V.A. Kozlov
"Candidate SNP-markers of rheumatoid arthritis that can significantly alter the affinity
of the TATA-binding protein for human gene promoters"

Supplementary 1. DNA sequence analysis

Two DNA sequence $S_{WT}=\{s_{WT;-90}\ldots s_{WT;i}\ldots s_{WT;-1}\}$ and $S_{SNP}=\{s_{SNP;-90}\ldots s_{SNP;i}\ldots s_{SNP;-1}\}$, which lengths are 90 bp that correspond to two variants of a given promoter located immediately upstream of the transcription start site (TSS, $s_{1;0}=s_{2;0}$; where $s_i \in \{a, c, g, t\}$) are the input data of the Web service SNP_TATA_Z-tester used (see the main section "Materials and Methods", Figure 1; URL=http://beehive.bionet.nsc.ru/cgi-bin/mgs/tatascan_fox/start.pl).

First of all, the affinity estimate "$-\ln(K_D(S))$" is calculated upon each of these sequences $S \in \{S_{WT}, S_{SNP}\}$, as:

$$-\ln(K_D) = 10.9 - 0.2\ \{\ln(K_{SLIDE}) + \ln(K_{STOP}) + \ln(K_{BEND})\}, \tag{1}$$

where 10.9 (ln units) and 0.2 corresponds to the estimates of nonspecific TBP-DNA affinity (i.e., 10 mM (Hahn et al., 1989)) and the stoichiometric coefficient (Ponomarenko et al., 2008); K_{STOP} as an empirical estimate of an impact of the TBP stops at the most probable TBP-binding site according to Bucher's rule (Bucher, 1990), namely:

$$\boldsymbol{ln(K_{STOP})} = \underset{(+),(-)\ \boldsymbol{DNA\ chains}}{\mathbf{MAX}} \left\{\sum\nolimits_{\boldsymbol{j=-1}}^{\mathbf{13}} \boldsymbol{w_{j;s_{i+j}}}\right\}; \tag{2}$$

where w_{js} as an element of the Bucher's position-weight matrix (Bucher, 1990), which corresponds to the case of the nucleotide s located within j-th position of the DNP sequence analyzed.

In Eq. (1), K_{SLIDE} as an empirical estimate of an impact of the TBP sliding along DNA near the most probable TBP-binding site mentioned above (i.e., DNA sequence region [TBP-DNA contact ± 5bp]) is heuristically calculated, as:

$$-\ln(K_{SLIDE}) = \mathrm{MEAN}_{[\text{TBP-DNA contact}\pm5\text{bp}]}\ \{0.8[TA] + 3.4\mu + 35.1\}, \tag{3}$$

where [TA] as as weighted number of dinucleotide TA; μ as the arithmetical mean of the minor groove width of the DNA helix (Karas et al., 1996) of the TBP-binding site under consideration; 0.8, 3.4, and 35.1 as regression coefficients (Suslov et al., 2010b).

In Eq. (1), K_{BEND} as an empirical estimate of an impact of the DNA helix bend stabilizing TBP-DNA complex is calculated, namely:

$$-\ln(K_{BEND}) = \mathrm{MEAN}_{\text{TBP-DNA}}\ \{0.9[TA, AA, TG, AG] + 2.5[TA, TC, TG] +14.4\}, \tag{4}$$

where $\mathrm{MEAN}_{\text{TBP-DNA}}$ as the arithmetical mean of both DNA strands of the TBP-DNA complex under consideration (see Eq. (2)); 0.9, 2.5, and 14.4 as regression coefficients (Suslov et al., 2010b);.

Additionally, the "$-\ln[K_D]$" values (Eq. 1) are accompanied by its standard deviation estimates (δ) according to all the possible nucleotide substitutions, $s_{\bullet;j}\rightarrow\xi$, at each position j of the above regions [TBP-DNA contact ±5bp], such as:

$$\delta(S_\bullet)=[(\Sigma_{1\leq i\leq 26}\Sigma_{\xi\in\{a,c,g,t\}}[\ln(K_D(\{s_{\bullet;i-13}\ldots\xi\ldots{\bullet}s_{\bullet;i+12}\})/K_D(\{s_{\bullet;i-13}\ldots s_{\bullet;i+j}\ldots s_{\bullet;i+12}\})^2])/(3*26)]^{1/2} \tag{5}$$

Finally, two estimates "$-\ln(K_D(S_{WT}))\pm\delta(S_{WT})$" and "$-\ln(K_D(S_{SNP}))\pm\delta(S_{SNP})$" calculated upon the input sequences S_{WT} and S_{SNP} (Eqs. (1–5)) were statistically compared with one another in the terms of Fisher's Z-score, such as:

$$Z = \mathrm{abs}[\ln(K_{WT;D}/K_{SNP;D})]/[\delta_{WT}{}^2+\delta_{SNP;}{}^2]^{1/2}. \qquad (6)$$

where Z as the above Z-score pinpointing *p*-value of the probability estimatee of acceptance of the H_0-hypothesis "H_0: $K_D(S_{WT}) \neq K_D(S_{SNP})$", which was taken from the commonly accepted statistical package R (Waardenberg et al., 2015)].

On this basis, the final decision is made at its statistically significant level $\alpha < 0.05$ (where $\alpha = 1$-p), namely:

```
IF {INEQUALITY "$-\ln(K_{WT;D}{}^) > -\ln(K_{SNP;D})$" is statistically significant},
THEN {DECISION is "$S_{SNP}$ provides an underexpression of a given gene in comparison with $S_{WT}$, which is the norm"};
ELSE [IF {INEQUALITY "$-\ln(K_{WT;D}{}^) < -\ln(K_{SNP;D}{}^)$" is statistically significant},
THEN {DECISION is "$S_{SNP}$ provides an overexpression of a given gene in comparison with $S_{WT}$, which is the norm"},]
OTHERWISE {DECISION is "alteration of the expression of this gene is insignificant"}.
```

One can see this DECISION in Figure 1, such as: the text box "Result" of the Web service SNP_TATA_Z-tester.

Supplementary 2. Key-word search in the PubMed database

The key-word search within the data base PubMed is handmade using the standard facilities of this data base in the cases of each candidate SNP-markers predicted, which is the initializing data of the algorithm shown in Figure S.

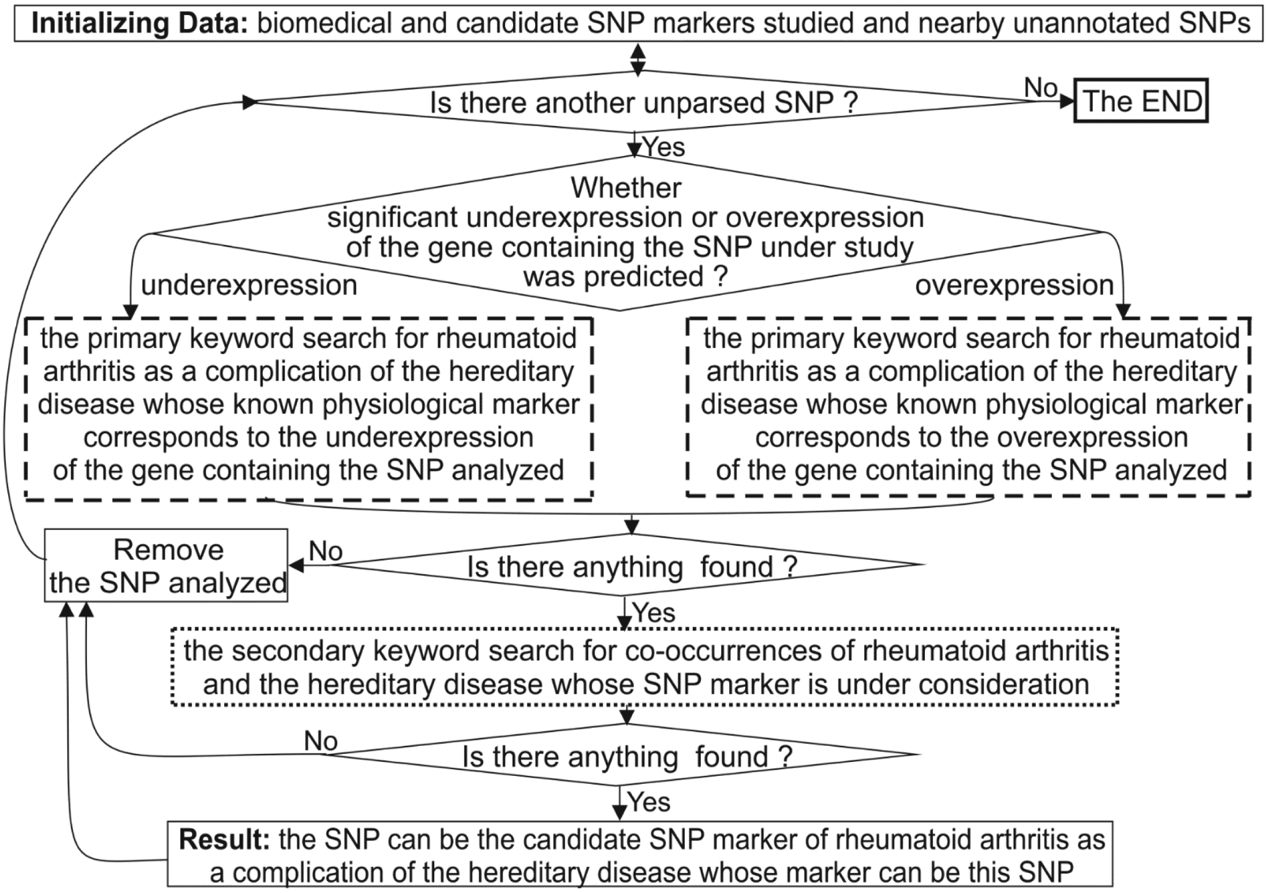
Initializing Data: biomedical and candidate SNP markers studied and nearby unannotated SNPs
Is there another unparsed SNP ?
No
The END
Yes
Whether significant underexpression or overexpression of the gene containing the SNP under study was predicted ?
underexpression
overexpression
the primary keyword search for rheumatoid arthritis as a complication of the hereditary disease whose known physiological marker corresponds to the underexpression of the gene containing the SNP analyzed
the primary keyword search for rheumatoid arthritis as a complication of the hereditary disease whose known physiological marker corresponds to the overexpression of the gene containing the SNP analyzed
Remove the SNP analyzed
No
Is there anything found ?
Yes
the secondary keyword search for co-occurrences of rheumatoid arthritis and the hereditary disease whose SNP marker is under consideration
No
Is there anything found ?
Yes
Result: the SNP can be the candidate SNP marker of rheumatoid arthritis as a complication of the hereditary disease whose marker can be this SNP

Figure S. A flow chart of the keyword search for rheumatoid arthritis (RA) as a comorbidity of hereditary diseases whose candidate SNP markers can alter TBP-binding sites in the human gene promoters.

In Figure S, two boxes (dashed lines) are corresponding to the primary key-word search for rheumatoid arthritis (RA) as a complication of the human hereditary diseases whose known physiological marker corresponds to the overexpression of the gene containing the predicted candidate SNP marker of RA, being under consideration.

As a sort of an independent non-statistical verification of the result obtained, one more (secondary) key-word search for co-occurrence of RA and the hereditary disease clinically associated with the gene containing the SNP being considered is additionally handmade by the same way (a box outlined with a dotted line).

Each positive outcome of these both independent keyword search steps following one another is the prediction made upon the SNP being tested as a candidate SNP marker of RA as a complication of the human hereditary disease verified successfully (see the main section “Results and Discussion”, Table 1) whereas the negative one is not (data not shown).