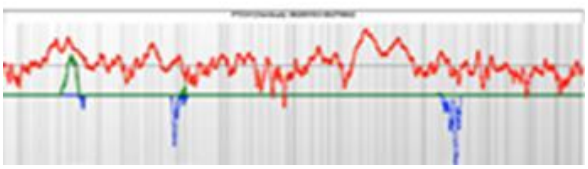
Entromics uses non-statistical approach to quantitative and personalized characterization of personal genome and individual genomic variants, found by modern experimental methods (microarray or whole-genome sequencing). This approach combines exact mathematics of information and coding theory, mathematical graph theory and statistical physics to obtain simple, easily manageable but highly informative and non-local characterization of the genomic variants or mutations. This novel approach has two levels of non-locality.

On the first level, we quantitatively characterize every position in the genome by the contribution of the base at that position to information entropy of the surrounding ±100-base up-stream/down-stream segment of the genomic DNA sequence. For reasons explained below, we call this descriptor the incorporation entropy (IE). This IE computation is done for wild type and variant DNA bases in a locus, leading to the characterization of the mutation by the difference of the two entropies. The computation is non-parametrical and fast, allowing whole-genome processing. This is the consequence of using the maximal entropy principle paradigm in the derivation of this descriptor. The resulting (analytical) formula takes the DNA sequence (represented by adjacency matrix of an Eulerian graph) of the 200-base gene segment, centered at a given locus of the wild type and variant gene and computes directly the IE's and their differences ΔIE = IE(var) – IE(wt). This computation scans systematically the whole chromosome and therefore the change of a single base results in the alteration of IE descriptors within ±100 base positions adjacent to the locus (see **Fig.1**



where red is IE(wt) for human PTCH1 gene and green-blue profiles are ΔIE = IE(var) – IE(wt) for three SNP's at different PTCH1 loci).

We therefore compute the positive, negative and total (squared) areas of resulting difference incorporation entropy profiles ΔIE and use these values as the final local descriptors of a variant/mutation in the central position. This integration of single nucleotide polymorphism impact on incorporation entropies over the entire local-context further increases the genome sequence-dependent information content of the descriptor, which – in turn – increases the likelihood of successful interpretation of the study results.

There are multiple advantages of this level of gene variant characterization:

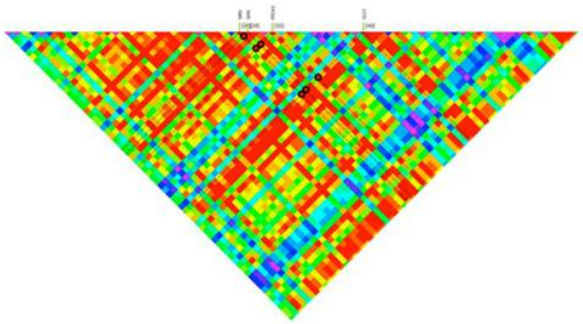
1. The base changes in two different loci have always non-identical ΔIE, even if the base changes involved are identical (with exception of long repeat regions, the 200-base DNA segments are all unique in genome).
2. The variant is characterized by a weight, a signed real-valued number, reflecting the local DNA sequence context of the variant locus. Substituting a broad spectrum of ΔIE values for the conventional categorical 0/0.5/1 allele descriptors definitely increases the sensitivity and information content of the "genetic independent variable", used in the interpretive models.
3. ΔIE is by definition either positive or negative. This sign provide natural, direct, a priori categorization of respective variants (SNP's) that can be tested for (functional) significance using the study results.
4. The IE is state function (its differences are independent of the path leading from the wild type to variant genome status) and therefore the IE values (and their differences and other linear combinations) computed in multi-loci studies for individual variants are additive. ΔIE computed for all individual variants can be therefore combined into cumulative descriptors without losing information. This can be used with advantage in genome-wide studies (dimensionality

reduction) or in functional modeling (variants are combined into separate descriptors according to the biological hypothesis, pathway etc.)

The second level of non-locality is again the direct consequence of deriving our genomic descriptors from the maximal entropy principle paradigm. We have guarantee that IE representation of the (local) biological information, carried by a base in the genome, is the best possible, given the 200-base DNA context of the locus. These local results say nothing about the optimality of this characterization in the context of the whole chromosome or genome. We have shown that for assembly of informationally optimal genome from these segments, the second application of the maximal entropy principle is necessary. This principle results in a unique way of combining the local genome segments into the chromosome sequence, so the local optimization of the information entropy extends to the complete genomic DNA. The uniqueness has the following important molecular consequence, distinguishing the biological; functionally and evolutionary optimized genomic DNA from a random sequence of the same set of bases: We found (with exact mathematical proof) that maximal entropy preserving assembly of genomic segments into genome requires that there are multiple segments with the same (or closely similar) local IE in the resulting genomic DNA.

In the next step, we have to explain, how the IE coherence is reflected in the molecular properties of genomic DNA. It can be directly shown that the two coherent DNA sequences, which by definition have identical incorporation entropy profiles, are not sequence-similar in conventional sense. At the same time, highly similar DNA sequences do share the incorporation entropy. So the IEC, a sequence-dependent (and therefore fundamentally molecular) property, generalizes the concept of sequence similarity. Consequently, IEC allows identifying larger numbers of long-range correlations in genome (and across the genomes), than can be found using the sequence similarity.

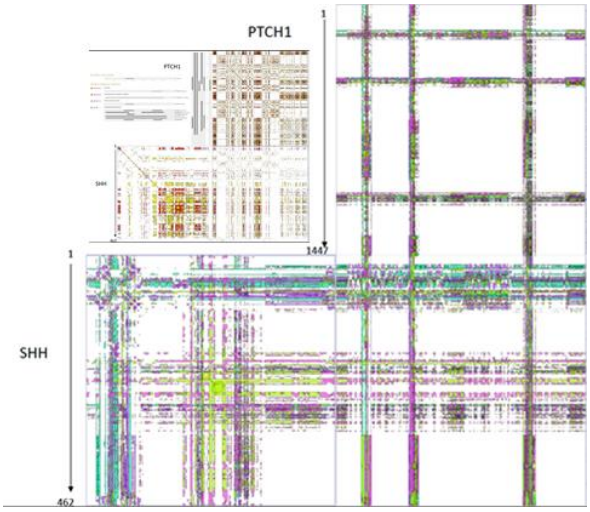
Because the IEC derivation is mathematically constructive, we can identify the IE coherences along any known genome DNA sequence. We do that by systematic comparisons of overlap integrals between vectors, computed from two series of sequence “windows” with increasing lengths, each centered in one locus of a pair. These oligomeric “windows” are increasing systematically, one base at a time, from ±10 to ±50 bases up- and down-stream from each locus. By definition, on the relative scale, these overlaps will be close to 100% for coherent pair of loci and close to 0% for two non-coherent genome regions. This information capacity maximization is general for any part of the genome. We can therefore compute the coherence matrices and their differences for “exome” DNA, defined as the coding DNA sequences of all genes, concatenated in the same order, as they are found along the chromosomal DNA.



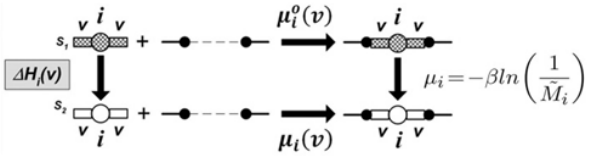
For practical purposes, we use the advantage that IEC values are also additive: The overlap integrals we compute to quantify IEC are constructed numerically as combinations of additive IE, therefore, IEC values are also additive. All relevant global coherence information can be summarized into fully informative, but simple descriptors. In our software platform, the IEC, represented as matrices, are computed from the reference (consensus) genome DNA (with major alleles or wt bases in all positions) and then form the

genomic DNA sequence with all personally genotyped variants incorporated. From these two IEC matrices, the difference IEC surfaces ΔIEC = IEC(var) – IEC(wt) are computed. To preserve numerical accuracy, this is done separately for genome regions with low and high local IE's (see Fig. 3).

Finally, to justify using the local and global incorporation entropy descriptors as input into the functional models of diseases, we needed to experimentally validate the link between the incorporation entropy and its coherence to measurable molecular properties of genomic DNA.



We have shown that DNA segments, which have coherent IE, have identical thermodynamic stability [1,2]. Similarly as in statistical thermodynamics, this observation reflects the known fundamental duality of entropy interpretation in both information and energy terms. With this result, the following thermodynamic cycle characterizes the complete difference in the genomic molecular properties between the wild type (top part of the scheme) and variant DNA (bottom part of the scheme).



If the DNA segments S_1 and S_2 are coherent, then they are thermodynamically iso-stable and the corresponding enthalpy difference ΔH is zero. Then these two positions are characterized by $\mu = IE$, which now has the natural meaning of the chemical potential μ of the reaction, incorporating the segment S_i into its position in genome.

This theoretical result is directly experimentally testable using the next-generation sequencing data: As a direct consequence of the maximal-entropy derivation of conditions for optimized chromosomal DNA assembly, we obtained the mathematical formula for the distribution function of the IE in the genome [#]. It is a function, formally identical to Planck distribution. The next-generation sequencing experiments [Pacific Biosciences] use enzymatic replication of the genomic DNA from fluorescently labeled nucleotide monomers. In the molecular reaction chamber, the detector measures not only fluorescence color of every base label that is incorporated into the replicated DNA strand by the immobilized single nuclease molecule, but also the time needed for that incorporation. This is the information relevant for our validation: the higher the activation barrier for the base incorporation, the longer is the time for that particular chemical reaction step and the longer the detected fluorescence "pulse" will be (the label is dissociated once the nucleotide is incorporated into the DNA polymer and is removed from the detector range by fast diffusion, so the pulse duration monitors the time the monomer spent in the reaction site, unattached to the replica). We

retrieved the fluorescence duration data from database of human genome sequencing results [Pacific Biosciences], constructed histograms of the experimental incorporation times and have shown that they have the (non-symmetrical) distribution function, theoretically predicted for IE.

We further tested this molecular interpretation of IE in the following way: Once IE is interpreted as the (sequence-dependent) chemical potential μ of the nucleotide incorporation reaction (see Fig.), then $\beta = 1/kT$, where k is the Boltzmann constant, T temperature. The same constant β is part of the derived IE distribution function formula. We therefore directly computed the IE for every position in hundreds of 1M base regions along the human genome DNA. We then least-square fitted the histograms of IE values from these regions by the theoretical distributions. In the final step, we confirmed that numerically correct value of k is extracted by processing the optimal parameters of the theoretical distribution function, obtained from these fits.

The biological consequence of this classical physical evidence for soundness of the underlying theory provides a guide for a functional interpretation of the IE and IEC. These descriptor values are proportional to the thermodynamics and reaction kinetics parameters, describing the gene expression, regulation and other processes, and influencing the system dynamics of the cell.

Thus having the physically correct, deterministic, non-statistical formulae, computing IE and the incorporation entropy coherence (IEC) from the personal genomic data converts this component of "high-throughput" characterization of a patient into highly relevant, a priori computed information in clinical applications.

The main biological significance of existence of IEC is in providing the quantitative, non-statistical measure of the impact of local base change in a single locus on the functionally relaxant context in all other positions of the processed sequence. As the numerical characteristics of this collective, subject specific, individual perturbation of the common, all-major allele state of the genome (or studied genes), we compute positive, negative and total (squared) volumes of the difference profiles ΔIEC along all studied regions. In particular, if genotyping data are available for multiple genes, then we compute these global ΔIEC differences not only for "diagonal" blocks of positions of respective individual genes, but also for the complete set of "off-diagonal" blocks, describing where, with what sign and with what intensity are the mutual variations impacting the long-distance communication between all studied genes.

REFERENCES:

1. Pancoska P, Moravek Z, Moll UM.: Nucleic Acids Res. 2004 Aug 27;32(15):4630-45. Rational design of DNA sequences for nanotechnology, microarrays and molecular computers using Eulerian graphs.,
2. Pancoska P, Moravek Z, Moll UM. Nucleic Acids Res. 2004 Mar 1;32(4):1469-79. Efficient RNA interference depends on global context of the target sequence: quantitative analysis of silencing efficiency using Eulerian graph representation of siRNA.