sciendo

Orvostudományi Értesítő
Bulletin of Medical Sciences
www.orvtudert.ro

Genomika és informatika – egy „szerelmi házasság" története

Falus András

Semmelweis Egyetem, Budapest

AUTHOR AFFILIATION

Department of Genetics,
Cell- and Immunobiology
Semmelweis University, Budapest

CORRESPONDING AUTHOR

Falus András
Professor Emeritus
Semmelweis Egyetem,
az MTA rendes tagja
falus.andras@med.semmelweis-univ.hu

DOI: 10.2478/orvtudert-2019-0013

Genomics and informatics The story of a „love marriage"

ABSTRACT

The rapid development of molecular biology over the past seventy years led to the molecular understanding of the structure and function of hereditary material in the living world. Through the processing of the gigantic amount of data obtained as a result of technological advances and the exploration of different mechanisms of action, information technology has become an indispensable part of genomic research. This article addresses the background of the genomic and bioinformatics revolution, its application areas, and the perspective of personalized medicine.

Keywords: genome, genomics, bioinformatics

KIVONAT

A molekuláris biológia gyors fejlődése az elmúlt hetven évben lehetővé tette az örökítő anyag szerkezetének és működésének molekuláris szintű megismerését az élővilágban. A technológiai fejlődés következtében rendelkezésre álló gigantikus adatmennyiség feldolgozása és a működési mechanizmusok feltárása nyomán az informatika elengedhetetlen részese lett a genomikai kutatásnak. A cikk a genomikai és bioinformatikai forradalom hátterével, egyes alkalmazási területeivel és a személyre szabott gyógyítás perspektívájával foglalkozik

Kulcsszavak: genom, genomika, bioinformatika

Genetika és genomika; a rész és az egész

Az emberi szervezetben mintegy százbillió (10^{14}) sejt található. Minden egyes sejtünk sejtmagjában 2x23 darabra vágott kromoszóma (ivarsejtekben a fele), 2x3,2 milliárd (10^9) négyféle nukleotidbázis (ezek: adenin - A, guanin - G, citozin - C és timin - T) található, ez a dupla helikális szerkezetben kb. 2x2 (4) méter DNS-t jelent. Az RNS timin helyett uracilt tartalmaz. A nukleotidbázisok lineáris sorrendje képezi a szüleinktől örökölt biológiai hardvert. Az élet során különböző hatásokra természetesen megváltozhatnak a nukleotidok (csere, kiesés, beékelődés, átrendeződés), ezeket a változásokat mutációknak nevezzük.

Az egyes génekkel a genetika, az összessel (beleértve azok kölcsönhatásával is) pedig a genomika foglalkozik. A genetika legfontosabb felfedezéseinek (a DNS mint örökítőanyag azonosítása, az öröklődés törvényeinek felismerése, a DNS szerkezetének leírása) sorába illik óriási továbblépésként a teljes human örökítőanyag (genom) szekvenciájának megállapítása.

A szekvenálás legközvetlenebb eredménye az összes fehérjét kódoló mintegy 23-25 ezer gén azonosítása. Az emberi gének viszonylag csekély száma (mely nagyságrendileg hasonló a fonalféregben találtakhoz!) rávilágított arra, hogy a biológiai fenotípus (tehát valóságos megjelenés) komplexitását nem a génkészlet nagysága, hanem magukban a gének variánsaiban rejlő egyedi sokféleség (diverzitás), a kapcsolati gén- és géntermék hálózatok szövevénye, valamint a gének megszólalására ható epigenetikai hatások sokasága határozza meg. Ez alatt nemcsak a fehérjekódoló gének működését szabályozó mechanizmusok összetettségét kell érteni, hanem a szabályozó rendszerek (pl. kis RNS-ek, metilációk, hisztonmódosítások, 3D kromatin átrendeződések) hálózatát is. A gének (fehérje és RNS gének) felsorolásán túlmenően ma már a gének pontos helyét és sorrendjét is nagymértékben ismerjük a genomban. Ez az egyszerű információ óriási jelentőségű a genetikában, mert lehetővé teszi azt, hogy egy kromoszóma szakaszhoz kapcsolt („térképezett") tulajdonsághoz vagy betegséghez gének módosulásait, variációit rendelhessünk hozzá.

A genomot tekintve csillagászati méretekről van szó, hiszen ha az összes emberi sejttel számolunk, az emberi szervezet DNS-hossza mintegy 140-szerese a Föld–Nap távolságnak.

Az emberiség DNS szinten is nagyon egységes, a rasszizmus minden álságos biológiai alátámasztása nemcsak morálisan elfogadhatatlan, hanem biológiailag, tudományosan is hamis. Az egyes etnikumok között néhány tizedszázalékos eltérés van a genom szintjén.

A humán genom program (1989–2003) lehetővé tette ennek a gigantikus információ „elolvasását". Ebben a 13-14 évben, rendkívüli nemzetközi együttműködéssel leírták a genomot alkotó mintegy 3,2 milliárd építőelem (nukleotidbázisok: A, C, G és T) lineáris sorrendjét. A két nagy konkurens, az államilag támogatott HUGO (Human Genome Organization) illetve a Celera privát cégből kinőtt magánvállalkozás természetesen csak kevés egyedi genomot tudott „elolvasni", ennek megfelelően messze nem volt világos, hogy mely genetikai „szavak és betűk" találhatóak meg minden emberben és melyek valóban egyediek.

Egy emberből átlagosan 20 fehérjekódoló gén teljesen hiányzik, azaz ebből a szempontból „génkiütöttnek" azaz KO-nak tekinthető. Ezek általában olyan gének, melyek hiánya nem okoz evolúciós hátrányt a ma élő embernek. Ilyenek pl. egyes szagreceptorok hiányai. Vannak olyan génhiányok viszont, amelyek kisebb hátrányt, vagy előnyt jelenthetnek hordozójuknak.

A humán genom mintegy 45%-a ismétlődő szekvenciákból áll. Ezek közül sok a transzpozon, azaz ugráló gén, amelyek viszont akár 40 millió év óta is inaktívak. A legygyakoribb ismétlődő szekvenciát Alu-nak hívják, mely a teljes genomunk 10,6%-át foglalja el.

Több száz génünk származik baktériumokból horizontális gén-transzferből.

A pericentromerikus és a subtelomerikus régiókban nagy szakaszok ismétlődnek

Jelenleg az imprintált gének számát 150 körülire becsülik (genetikai imprinting: az eltérő apai és az anyai gének kifejeződése), amelyek közül vagy csak az anyai (56%), vagy csak az apai (44%) aktív, de a pontos számok vitatottak. Ha valami oknál fogva ebben a rendszerben hiba következik be, tehát pl. ha mindkét gén aktív, súlyos betegségekhez vezet (pl. Beckwith-Wiedemann és Angelman szindrómák).

A CpG szigetek olyan 200 bázispárnál (bp) hosszabb szekvenciák ahol a CG dinukleotid arány magasabb a vártnál. Ezekből 27.000-29.000 db található az ismétlődésmentes részeken; sokszor egybeesnek a gének 5' végével (40%). A citozinon metilálódhatnak, amivel befolyásolhatják a gének expresszióját, szerepet játszanak a gén inaktivációjában és az imprintingben. Általában a promóter régió metilációja a transzkripciós aktivitás csökkenését, a kódoló régió metilációja a növelését okozza. A metilációs mintázat erősen sejtspecifikus. Az őssejteken a metiláció 25% nem CG-n történik, hanem CA-n (szemben a normál sejtekkel, ahol ez az arány csak 1%).

Eddig kb. 15.000 pszeudogént találtak. Ezek inaktív gének: lehetnek nem expresszálódó másolatok: processed (intron nélküli), unprocessed duplicated (intronos) változata az eredeti génnek; de átíródhatnak RNS-sé is. Korábban semmilyen szerepet nem tulajdonítottak nekik, azonban újabb kutatások alapján, az átíródó pszeudogének befolyásolhatják a velük rokon gének működését, pl. úgy, hogy kompetícióba kerülhetnek a génexpresszió szabályozásban fontos szerepet betöltő miRNS-ekkel, vagy expressziójukkal csökkenthetik a funkcionális gén stabilitását. Becslések alapján a pszeudogének 9%-a íródik át.

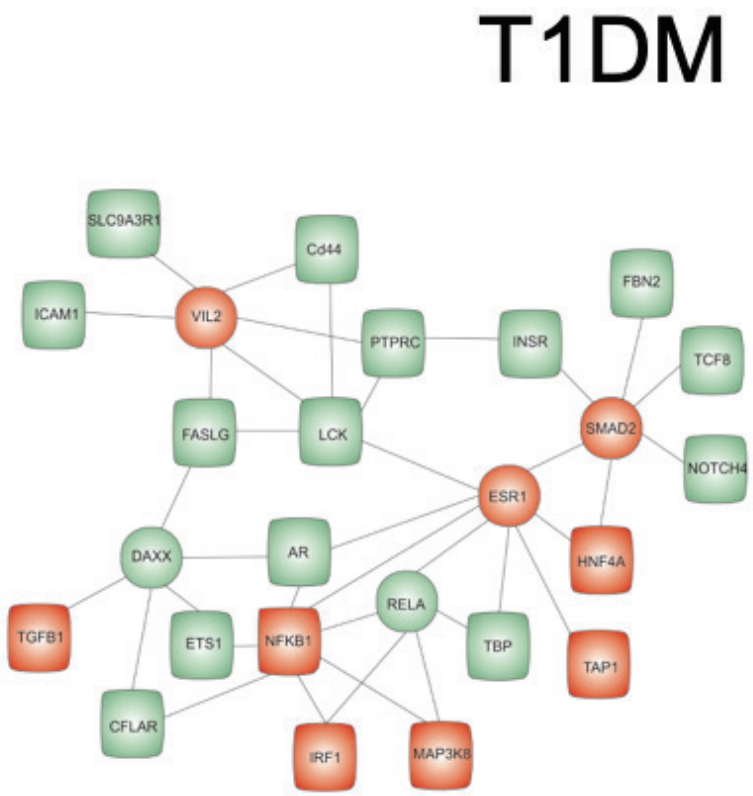
A human genom szekvencia sikeres leírása, első „munkapéldánya” (mint „draft”) pontatlanságai ellenére is vitathatatlan mérföldkő volt a genetikában, hiszen a genomszekvencia a genetika olyan alapdokumentumává vált, ami nélkül a genetikai tudományok további fejlődése elképzelhetetlen volt. Olyan ez, mint egy könyv szövege, betű- és szóhalmaza, ami szükséges (de nem elégséges) feltétel a „szöveg” megértéséhez. Önmagában ezzel az „írásjeltömeggel” még nem tudunk mit kezdeni, a nyelv, a biológiai „nyelvtan” ismerete nékül csak értelmetlen ákombákomnak látjuk.

A genomika és az informatika kapcsolata

Az utóbbi években egyre több és szabadon (puszta regisztrációval) elérhető adatbázis vált hozzáférhetővé a kutatók számára. Létrejött a térbeli és időbeli korlátokat virtuálissá tevő „in silico” (komputer előtti) kutatás lehetősége. Ez egyben a tudomány rendkívül széleskörű demokratizálódásával járt, hiszen bárki a világon könnyen és legtöbb esetben ingyen felkeresheti ezeket az adatbázisokat interneten. Ezt követően saját hagyományos kutatólaboratórium („nedves labor”) nélkül is a meglevő adatok új megközelítésével, csoportosításával, csupán a számítógép mellett önálló és eredeti tudományos felfedezéseket tehet. Ez a helyzet tudománytörténeti jelentőségű, kölcsönösen megtermékenyítő az informatika és a biológiai tudományok (pl. orvosbiológia) számára.

Szükségszerűen kiteljesednek a bioinformatikai elemzések a nagy elemszámú biológiai rendszerek adattengerének elemzésére is. Napjainkban útvonal- és génhálózat analízisek és az ennek megfelelő szoftverek sokasága jelent és jelenik meg. Ezek a megoldások teljesen új felfedezési stratégiákat alapoztak meg a preszimptomatikus prevencióban, illetve betegségek gyorsabb felismerésében és gyógyításában is.

Példaként, az 1-es típusú (inzulinfüggő) cukorbetegség ismert génjei alapján, hálózatelvi alapon 68 (!) új, eddig nem ismert, illetve nem azonosított gént találtak. Ezek (pl. ESR1, VIL2, SMAD2, RELA, DAXX) jelentősen hozzájárultak a diagnosztika érzékenyebbé és pontosabbá válásához, illetve potenciális terápiás célpontot jelenthetnek a gyógyszerfejlesztés számára. Az **1. ábrán** a teljes génhálózat egy részét mutatjuk be (zöld színnel az eddig ismert gének, sárgával a hálózati struktúra által feltárt „új”, azaz újonnan felismert gének).



1. ábra. A teljes génhálozat egy részlete

Ez a példa utal az ún. „preciziós medicina" új lehetőségeire is, tehát a személyre szabott diagnosztika és terápia előretörésére.

Új biostatisztikai korszak

A hagyományosabb, ún. „frekventista" analízisek mellé beléptek a nagy halmazokat kezelő matematikai–statisztikai eljárások. Ezek, pl. a BN-BMLA (Bayesian multilevel analysis) random változók közötti kapcsolatok valószínűségének eloszlását mutatja.

Nem véletlen, hogy ma már a molekuláris- és genomszintű vizsgálatok anyagi feltételei közül a szuperszámítógépek, a folyamatosan megújított szoftverek és az azokat fejlesztő jól felkészített informatikusok (bioinformatikusok) tudása minősíthető az egyik legkeresettebb (és jól fizetett) hivatásnak. Ez a területek kiemelt lehetőségeket adnak a mesterséges intelligenciák, a nanobiotechnológia és a robotika számára is.

A modern genetika tehát ma már elválaszthatatlanul kapcsolódik az informatika tudományához.

Ma már az ún. „posztgenomikus" korban élünk, a lexikális megismerésen túl a működés, szabályozás, és a gének funkcióinak feltárása, az „annotáció" zajlik. Megtudtuk, hogy az örökítő anyag óriási elemszámú hálózatokban működik. A teljes rendszer áttekintését célzó megközelítésre szolgál a rendszerbiológia vagy rendszer-szemléletű biológia (systems biology) elnevezés, amely egy teljesen új „csapat-függő" világot nyitott meg a kutatók, orvosok, biotechnológusok, matematikusok számára.

Nyilvánvaló, hogy tudásunk validálásához még sokkal több ember genom szekvenciájának megismerésére lesz szükség.

2012-ben fejeződött be az ún. „1000 genom projekt", ennek alapján jött létre az ENCODE, ami egy genetikai enciklopédiának felel meg. Kínai genetikusok közeli célul tűzték ki több millió ember teljes genomjának elolvasását. A viharosan fejlődő módszerek, például az új generációs szekvenálási eljárások, és a rohamosan csökkenő költségek folytán valószínű, hogy ez a cél pár éven belül meg fog valósulni.

A továbbiakban meg kell tudnunk mondani minden egyes variánsról, hogy hozzájárul-e a betegséghez, vagy például egy adott gyógyszer lebontásának kinetikájához, s ha igen, milyen mértékben. Ennek megállapítása igen nehéz feladatnak ígérkezik, tekintve, hogy a betegségeket okozó variánsok száma valószínűleg igen nagy és a legtöbb etnikumban, sőt egyes emberekben is különböző.

Mindazonáltal ennek a genetikai információnak a birtokában prediktív módon megbecsülhető lesz majd a betegségek kialakulásának genetikai kockázata még azok bekövetkezte előtt.

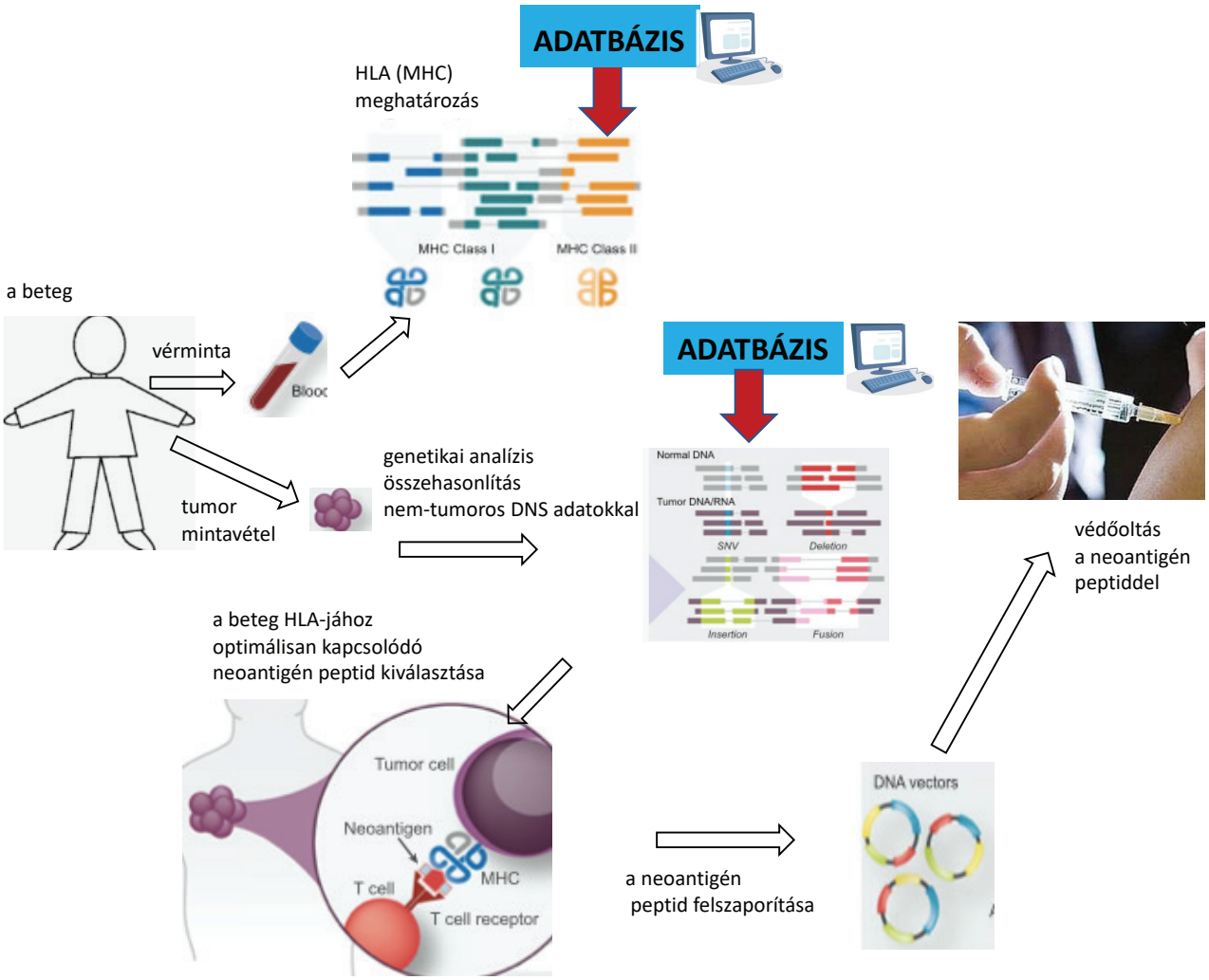
Az informatikai analízisek egyre inkább a mesterséges intelligenciák alapvető felhasználása felé mutatnak.

Az informatikai analízis sémáját a rákkutatás egy példáján keresztül mutatjuk be. A tumorokra jellemző neoantigének kimutatása és aminosav sorrendjének meghatározása az egyik első kulcseleme a rákkutatásnak. Ma már ez a vizsgálat bioinformatikai jellemzések sorozatán keresztül valósul meg. A neoantigén jellemzésére és a tumorvakcina bioinformatikai előállítására szolgáló átfogó munkafolyamat főbb elemzési lépéseit egyszerűsített formában a **2. ábrán** mutatjuk be.

Először betegek örökölt immungenetikai sajátosságait, a fő hisztokompatibilitási fehérjéit MHC (emberben humán leukocita antigén: HLA) típusokat határozzák meg, felhasználva a nagy adatbázisokban talált információkat. Ezután a tumor genetikai analízise következik, szomatikus variánsokat keresnek a betegből izolált (biopszia, műtéti minta) tumorszövetben (pl. egy nukleotid cseréket, deléciókat, inszerciókat és fúziókat). Ezt követően az MHC fehérjékhez kapcsolódó neoantigén-peptidek közül informatikai („in silico") predikciót hajtanak végre az interneten hozzáférhető adatbázisok felhasználásával, azaz a beteg HLA molekuláira „illesztve" tervezik meg a legjobban kapcsolódó neoantigén eredetű peptideket. A kiválasztott peptideket ezután megfelelő hordozókkal (pl. vírusok) felszaporítják és vakcinákat állítanak elő. A vakcinák stimulálják a beteg immunrendszerét és immunológiai védelmet nyújtanak a daganat ellen. A kutatások nyomán egyre hatékonyabb tumorellenes védőoltások előállításával reményteli módon fel lehet venni a harcot a molekulárisan jellemzett daganatok ellen.

A „geneticizmus" veszélye-génetikai kihívások

A genetikai kutatás felgyorsulása alapvető jogi, etikai és világnézeti kérdéseket is felvet. Egyes genetikai adottságok öröklődése nem csak a családi, hanem etnikai összefüggéssel is rendelkezik, amit a halmozódásuk jelez a meghatározott társadalmi csoportokban. Ez együtt járhat a társadalmi stigmatizáció, a diszkrimináció, a kirekesztés jelenségeivel.



2. ábra. A neoantigén jellemzésére és a tumorvakcina bioinformatikai előállítására szolgáló munkafolyamat főbb lépései

A genetikai kutatás eredményeinek egyik elsődleges alkalmazási területe az emberi diagnosztika és gyógyítás. A bekövetkezett fejlődés hatására megváltozott a klinikai orvosi szemlélet és gyakorlat. Ennek súlypontja a tünetekkel jelentkező beteg kezeléséről fokozatosan a tünetmentes állapot idején folytatott prediktív diagnosztikai tevékenységre alapozott megelőzésre, valamint a személyre szabottan tervezett és végrehajtott orvosi beavatkozásokra helyeződik át. A betegjogi kérdések ma már teljesen új megvilágításba kerülnek, a genetikai információ különleges volta miatt.

A genetikai/genomikai/epigenetikai „hype“ (csinnadratta), a genetika kizárólagos jelentőségének túlhangsúlyozása („geneticizmus”) nagy veszélyt is jelenthet, mert a megismerés, a tudásunk és a gyakorlati hasznosíthatóság jelen fázisa kezdetinek tekinthető. A nagy hírverésel nyilvánosságra hozott ENCODE eredményei pár éve elmaradtak a várakozástól. Megjelent egy elég szkeptikus kifejezés, a hiányzó örökletesség (missing heritability). Ez persze nem az eredményeket, hanem a túlzóan (de talán érthetően) nagy elvárásokat minősíti. A genetika/genomika értelmezése markánsan eltávolodik a „sors” fogalmától, ma már e tudományok legfontosabb szavainak egyike a hajlam, amiben a valószínűség fogalmát is bele kell értenünk.

A hálózati gondolkodás mellett a külső és belső környezet által létrejövő, reverzibilis epigenetikai hatások figyelembevételével sokkal jobban helyére kerülnek a genetika és a genomika tudományának óriási és egyben valós eredményei és társadalmi hasznossága.

Irodalom

1. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml 2009.
2. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
3. Venter JC és mtsai The sequence of the Human Genome. Science 2001;291:1304-51.
4. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome Nature 431, 931 - 945 (21 October 2004)
5. http://genomics.xprize.org/
6. http://www.genome.gov/10005107; 2009.
7. Pennisi E. 1000 Genomes Project Gives New Map Of Genetic Diversity. Science 2010; 330: 574-5.)
8. http://www.epigenome.org/; 2009.
9. Redon R. és mtsai.:Global variation in copy number in the human genome. Nature2006; 444: 444-454.
10. Armour JA. Copy number variation and antigenic repertoire. Nat Genet. 2009;41(12):1263-4.
11. Bruder CE, és mtsai.: Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet. 2008;82:763-71.
12. Burbano HA, és mtsai Targeted investigation of the Neandertal genome by array-based sequence capture. Science. 2010 May 7;328(5979):723-5.
13. Gibbs W .W . (2003) "The unseen genome: gems among the junk", Scientific American, 289(5): 46-53.
14. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447:799- 816.