1. A genomika alapjai - A humán genom.

1.1. *Genomika*

Bár a genomika tudománya már több évtizedes múltra tekint vissza, tulajdonképpen csak az elmúlt két évtizedben vált még az élő természettudományokkal foglalkozók között is igazán ismertté. Annak ellenére azonban, hogy jelenleg is a leggyorsabban fejlődő tudományágak közé tartozik, a hétköznapi emberek túlnyomó többsége, sőt például a régebben végzett orvosok, gyógyszerészek számára is gyakorlatilag ismeretlen fogalmat takar. Éppen ezért a bevezetőben néhány fogalmat definiálok.

Először is: Mi az a genom? A **genom**: Egy diploid sejt teljes haploid DNS tartalma, plusz a mitokondriális DNS. Mivel a férfiak és a nők genomja különbözik abban, hogy a férfiaknak kétféle nemi kromoszómájuk van (X és Y), a genom meghatározásnál ezt is figyelembe kell venni. A következő fontos kérdés, hogy mivel foglalkozik a genomika? Igazából erre többféle definíciót is lehet adni, ezek közül talán a legegyszerűbb: A genom működésének, szerkezetének, kölcsönhatásainak vizsgálata és az ezekhez tartozó módszerek. A DNS vizsgálatán túl azonban ide tartoznak az RNS-ek vizsgálatai (transzkriptomika), a fehérjék vizsgálatai (proteomika), bioinformatika, rendszerbiológia is. Egyes meghatározásokban a genomikát a molekuláris rendszerbiológia szinonimájának is definiálják, amelyben a genom szintű szerveződések alapján ismerjük meg a világot. Valójában a genomika inkább a rendszerbiológia részének tekinthető. A genomika vizsgálatának tárgya alapján különböző alcsoportokra osztható. Lehet például: strukturális genomika; komparatív genomika; funkcionális genomika; humán genomika; farmakogenomika; orvosi genomika, stb. Ebben a könyvben főleg az utolsó hárommal foglalkozom.

Van még egy fontos kérdés, ami sokak számára problémát jelent. Mi a **különbség a genetika és a genomika között**? Igazából nem húzható éles határvonal a két tudományág közé, mindenesetre általánosságban elmondható, hogy ha egy gént, vagy genetikai variációt vizsgálunk, akkor genetikáról szoktunk beszélni, ha több gént, vagy az egész genomot mint rendszert vizsgáljuk akkor genomikáról beszélünk. Ebből következik, hogy a genomika, mivel a rendszerbiológia témakörbe tartozik, általában jóval összetettebb módszereket igényel. Azonban, még a tudományos szóhasználatban is, a két szó használata erős átfedést mutat.

1.2. Humán Genom Projekt

A genomika tudomány ugrásszerű fejlődését a Humán Genom Projektnek (**HGP**) köszönhette. A következőkben a teljesség igénye nélkül röviden ismertetem a HGP történetét, célkitűzéseit és néhány eredményét (1).

Érdekes módon a HGP az amerikai Energiaiügyi Minisztérium (Department of Energy, DOE) kezdeményezésére indult el. Ennek a minisztériumnak az elődjei voltak ugyanis az atombomba kifejlesztésének irányítói. Miután Japánban az amerikaiak ledobták a két

atombombát, az amerikai kongresszus azzal bízta meg a DOE elődjét, hogy tanulmányozza a genomot, hiszen a nukleáris sugárzás hosszú távú károsító hatásának az oka a genom sérülése. Mivel a genomot úgy lehet legjobban tanulmányozni, ha megismerjük annak felépítését, 1986-ban a DOE az NIH-el (National Institute of Health) összefogva elhatározta a HGP elindítását. A szervezőmunka 4 évig tartott, és **1990. október 1-én hivatalosan is elindult a HGP**.

A projektre 15 évet és **3 milliárd dollárt** szántak. A projekt **fő céljai** a következők voltak: Azonosítani a kb. 100.000 gént a humán genomban; Meghatározni a humán genomot felépítő 3 milliárd bázist; Nyilvános adatbázisokban tárolni az információkat, és szoftvereket fejleszteni az elemzéshez; Bevonni a magánszektort; Etikai, törvényi és társadalmi problémákat tisztázni; Modell szervezetek megszekvenálása (pl. egér, csimpánz, háziállatok; növények, mikroorganizmusok, patogének stb.). Ez utóbbival választ kaphatunk olyan kérdésekre mint pl: Miért ember az ember (csimpánz vs. ember)? Melyek az élethez nélkülözhetetlen gének (konzervált gének, amelyek minden élőlényben megtalálhatók)? Patogének, háziállatok, növények megszekvenálása.

A HGP nem ment teljesen zökkenőmentesen. 1998-ban, a projekt tervezett idejének a felénél, még csak az emberi genom 5%-a volt ismert, és reménytelennek látszott a projekt sikeres teljesítése. **Craig Venter** a projekt egyik vezető alakja egy új módszer bevezetését javasolta. Ez az ún. „shotgun sequencing" lényege az volt, hogy a genomot rövid darabokra felszabdalták, ezeket a darabokat megszekvenálták, majd az átfedő szekvenciák segítségével összeillesztették őket. A módszer bizonyítottan működött kisebb (bakteriális) genomoknál, azonban a HGP vezetői úgy gondolták, hogy a nagyságrendekkel nagyobb emberi genomnál ez a módszer nem fog működni. Venter ezért kilépett a HGP-ből és saját céget (**Celera**) alapított, és magántőke bevonásával be akarta bizonyítani, hogy jól működik a módszere. Ez a kiválás katalizátorként hatott az eseményekre. A gyorsuláshoz persze az is hozzájárult, hogy a 8 év alatt rengeteg olyan fejlesztés történt, amelyek hatása ekkora érett meg. A szekvenálást a Sanger által kifejlesztett didedoxi módszerrel végezték, amelynek egyik hátránya az volt, hogy a DNS-t felépítő négy nukleotid sorrendjét 4 külön zajló reakció segítségével állapították meg, melyek termékeit egymás mellett 4 külön csíkban kellett elektroforézis segítségével futtatni. Ezzel kapcsolatban kifejlesztették, hogyha 4 féle fluorescens festékkel jelölik meg a 4 terméket, azok egyszerre, egyetlen kapilláris elfo segítségével is futtathatók. Ráadásul, pl. az Applied Biosystem olyan szekvenáló automatát is kifejlesztett, amely egyszerre 96 kapillárissal tudott dolgozni és mindössze napi 15 perc beavatkozást igényelt. De kifejlesztették például a bakteriális mesterséges kromoszóma vektort, amelyekbe a korábbiaknál lényegesen nagyobb (100-200 kilobázis) genom darabot lehetett klónozni. Az egyéb jelentős fejlesztések mellett kiemelkedik a számítástechnika fejlődése. A humán genomot felépítő több mint 3 milliárd „betű" tárolása, kezelése, annotálása a 90-es évek elején még méregdrága szuperkomputereket igényelt, manapság már bármelyik személyi számítógép is könnyedén megbirkózik a feladattal. Ezzel kapcsolatban rengeteg új bioinformatikai módszer került kifejlesztésre és számos nyilvános, felhasználóbarát, on-line adatbázist alakítottak ki. Mindezek hatására 1999-től 15 hónap alatt 10%-ról 94%-ra nőtt a megismert humán szekvencia aránya. A szekvenálásokat nagy, gyárszerű épületekben végezték. A szekvenálás sebességére jellemző, hogy 1999-ben a HGP havonta 7 millió mintát dolgozott fel, és másodpercenként 1000 nukleotidot szekvenált meg. A Celera teljesítménye még lenyűgözőbb. A mindössze 65 fős személyzetből álló szervezet 1999 szeptember 8-a és 2000 június 17-e között 14.9 milliárd bázist szekvenált meg, ami a teljes humán genom közel 5-szörös lefedését tette lehetővé. A Celera és a HGP végül egymással megegyezve, egyszerre, 2001 februárjában, ugyanazon a héten közölték eredményeiket a világ két vezető tudományos lapjában a Nature-ben és a Science-ben (2,3). A közölt eredmények még csak a humán genom ún. „draft" szekvenciáját tartalmazták, azaz számos lyuk (gap) volt még benne, illetve sok

szekvenálási hibát tartalmazott. Ebben egy szekvencia 4-5-szörös lefedéssel volt megszekvenálva. A **HGP hivatalosan 2003 áprilisában z**árult, amikorra elkészült a lyuk nélküli, magas minőségű, 8-9-szeres lefedettségű humán genom szekvenciája (4).

1.3. DNS szekvenálás

A DNS szekvenálás fejlődése a HGP befejezése óta is töretlen. A humán genom szekvenciáját a HGP-ben Sanger módszerével a didedoxi szevenálással határozták meg. A HGP összköltsége 3 milliárd $ volt, Craig Venter genomjának megszekvenálása összesen 70 millió $-ba került. 2001-ben egy humán genom megszekvenálása minimum egy évbe tellett. Nyilvánvaló ekkora költség és ennyi idő nem alkalmas arra, hogy a genom szekvenálás a mindennapi rutinná váljon, de még arra sem, hogy sok ember genomját megismerjük, összehasonlítsuk. Az is lassan világossá vált, hogy ennek a módszernek a továbbfejlesztésével nem lehet jelentősen csökkenteni a költségeket, illetve az időt. Azonban a szakértők számára nyilvánvaló volt, hogy ha olcsóvá és gyorssá lehetne tenni a szekvenálást, az óriási előrelépést jelentene például a gyógyszerkutatásban, vagy a személyre szabott gyógyászatban, de a felhasználási lehetőségek száma szintén végtelen. Hogy a fejlesztéseket elősegítsék, létrehozták az **Archon X díjat** a genomikáért (1. ábra) (5). Ezt a díjat, 10 millió $-ral együtt az a cég kapja, amelyik képes 100 emberi genomot, maximum 10 nap alatt, maximum 1 hibával 100 ezer bázisonként, és nem többért, mint 10 ezer $/genom költségért megszekvenálni. Ez a felhívás 2006-ban sokak számára még kicsit utópisztikus vágyálomnak tűnt, azonban hamarosan kiderült, hogy a célt valószínűleg hamarabb fogják elérni, mint ahogy azt sejtették. Abban mindenki egyetértett, hogy nem is a kitűzött 10 millió $ az igazi tét, hiszen, ha egy cég jelentős piaci előnyhöz jut ezen a piacon, az ennél nagyságrendekkel nagyobb hasznot fog elérni. Mindenesetre számos cég, illetve szervezet egymással párhuzamosan olyan radikálisan új fejlesztéseket, újításokba fogott, hogy rövidesen többen is a cél közelébe kerültek. A didedoxi szekvenálás helyett olyan módszereket fejlesztettek ki, mint pl. a szekvenálás ligálással (Solid technológia, Applied Biosystem), a piroszekvenálás (454 technológia, Roche), vagy a szekvenálás reverzibilis terminátorral (Solexa technológia, Illumina). A módszerek annyira sikeresek voltak, hogy a szekvenálást 2007-ben megválasztották az év módszerének (6). Az új módszerek közül először 2007-ben a 454 technológiával James Watson genomját szekvenálták meg 2 hónap alatt 1 millió $-ért, ami még ugyan messze volt a kitűzött céltól, de máris óriási előrelépést jelentett a korábbiakhoz képest. Sőt az összes cég folyamatos fejlesztésben van, és egyre lejjebb szorítják mind az időt, mind a költségeket. Például az Applied Biosystem 2008-ban 60 ezer $-ért szekvenált meg egy genomot. 2011-ben az Illumina HiSeq 2000 rendszerével 30x-os lefedettséggel két emberi genom megszekvenálása 8 nap alatt történik, 6 ezer $ dollár/genom költséggel. A Complete Genomics nevű cég egy cikkében 40x-es lefedettséggel már 1.500 $-ról írt. De a verseny tovább folyik, új technikákkal, vagy a régiek tökéletesítésével szinte havonta jelennek meg újabb hírek a szekvenálás árának csökkenéséről és sebességének növekedéséről (2. ábra).
A szekvenálás fejlődését kihasználva olyan projektek indultak, mint pl. az 1000 genom projekt, amely különböző etnikumú populációkhoz tartozó, összesen 2500 ember megszekvenálását tervezi. A projekt 2008-ban indult és első eredményeit már 2010 októberében közölték (7,8). De ilyen projektek például a **Genome 10k projekt**, amely 10 ezer gerinces faj megszekvenálását tervezi (http://genome10k.soe.ucsc.edu/), vagy a 2011. márciusában indult **i5k projekt**, amely 5000 rovar megszekvenálását tűzte ki célul http://www.arthropodgenomes.org/wiki/i5K).

1.4. Résztvevők a humán genom projektben

Először is érdekes kérdés, hogy kiket szekvenáltak meg először? A Celera a Science-ben megjelent cikkében ezt írta: 21 kevert etnikumú önkéntes donort választottak ki (kor, nem önmeghatározott etnikum). Mindenkitől vettek 130-130 ml vért, a férfiaktól 5 adag sperma 6 hét alatt. Végül a minták minősége és az etnikumok diverzitását figyelembe véve 2 férfit és 3 nőt választottak ki: 1 afrikait, 1 kínait, 1 spanyol-mexikóit és 2 kaukázusit (3). A hivatalos HGP 2 centrumban gyűjtötte a donorokat, hasonló elvek alapján.
Érdekesség azonban, hogy később kiderült, hogy a szépen hangzó szempontokkal szemben a Celera végül Craig Venter genomját szekvenálta meg.
A HGP-ben összesen 18 ország vett részt, de közülük messze kiemelkedett az USA. A nemzetközi genom projektet a **HUGO** (Human Genome Organization) koordinálta. A HGP-ről és eredményeiről részletesen olvashatunk a HGP hivatalos honlapján (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml), illetve számos más, az interneten fellelhető forrásból pl.: http://en.wikipedia.org/wiki/Human_Genome_Project.

1.5. A HGP néhány eredménye

A HGP/Celera számos érdekes, sokszor váratlan eredményt hozott, amelyeket a tudományos világ azóta is folyamatosan frissít, illetve bővít. Talán a legmeglepőbb eredmény az volt, hogy a várt, 100 ezres nagyságrendű génszám helyett **alig több mint 20 ezer gént tartalmaz** a humán genom Az 1. táblázatban látható néhány statisztikai adat a humán genomról.

Néhány érdekesebb eredményt az alábbiakban láthatók, kiegészítve/kijavítva újabb eredményekkel, pl. az 1000 genom projektből:

- Legnagyobb gén: Dystrophin (DMD): 2,2Mb (2.221.182 bp)
- Leghosszabb kódoló szekvencia: titin: 104.076bp, < 34.500 aminosav)
- Leghosszabb exon: titin: 17.106bp
- Legtöbb exon titin gén: 351 db
- 20%-a a genomnak (605 Mb) génsivatag = >500 kb gén nélkül
- Géngazdag kromoszómák: 17,19,22
- **Leggazdagabb: 19-es:** 63,8Mb = 1450 gén (Ensembl): 22,4 gén/Mb
- Génszegény kromoszómák: 4,13,18,X, Y
- **Y a leggénszegényebb** összesen 62 gén; 1,0 gén/Mb
- Az intronoknak a 98,12% GT bázisok vannak az 5’ végén és AG 3’ végén; 0,76% GC-AG
- Rekombináció magasabb a nőkben, mint a férfiakban, de a mutációk száma magasabb a férfi meiózisban, azaz az emberekben található öröklődő mutációk többsége a férfiakban keletkezik.
- Minden újszülött 60 új mutációt kap szüleitől.
- Minden ember átlagosan 250-300 funkcióvesztése mutációval rendelkezik az ismert (annotált) génekben, melyek közül 50-100 olyan gén, amely valamilyen öröklődő betegségben szerepet játszik. Ez többek között jelzi, hogy miért veszélyes, ha egymással rokonságban élő párnak gyermeke születik. Ilyenkor nagy az esély, hogy a szülőpárban heterozigóta formában jelenlévő, recesszív betegség a gyermekekben megjelenjen. Illetve, a funkcióvesztéses mutációnál, mivel egyes létfontosságú fehérjék kisebb mennyiségben vannak jelen, megnövelheti bizonyos betegségekre való hajlamot, vagy legalábbis befolyásolja a hordozók fenotípusát. Érdekes megjegyezni, azonban, hogy ennek ellentétes hatása is lehet, amit majd a gén-környezet

kölcsönhatás részben tárgyalunk, azaz bizonyos környezeti tényezőkkel szemben (pl. fertőzés) növelheti az ellenálló képességet.

- A humán genom 46%-a ismétlődő szekvenciákból áll. Ezek közül sok a transzpozon, azaz ugráló gén, amelyek viszont kb. 40 millió év óta inaktívak. A **leggyakoribb 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 156 imprintált gént (**Genetikai imprinting**: az apai és az anyai gének kifejeződése különböző) ismerünk, azaz ezek közül vagy csak az anyai (56%), vagy csak az apai (44%) aktív. 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).
- CpG szigetek olyan 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%). Metilálálódhatnak, amivel befolyásolhatják a gének expresszióját, szerepet játszanak a gén inaktivációjában és az imprintingben. Viszont őssejteken a metiláció 25% nem CG-n történik, hanem CA-n (szemben a normál sejtekkel, ahol ez az arány csak 1%).
- Az AT-gazdag régiók génszegények
- Detektáltak 298 db paralógot, egy exonos gént (processed paralog), ebből 97 igazoltan működik. **Paralóg** = génduplikáció eredménye, működnek, intronos vagy intronnélküli változat, funkciója lehet ugyanaz, vagy hasonló, de más is mint az eredeti génnek (vs. ortológ). Mivel a szelekciós nyomás a duplikálódott génen kisebb, vagy hiányozhat, szabadon mutálódhat, így nyerve új funkciókat.
- Eddig (2011. november) 14.266 **pszeudogént** találtak. Ezek, szemben a paralogokkal inaktív gének: lehetnek nem expresszálódó másolatok: processed (intronnélküli), 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 kompetícióba kerülhetnek a gén expresszió szabályozásban fontos szerepet betöltő miRNS-ekkel, így befolyásolhatják a velük rokon gének működését.
- A génexpresszió szabályozásában fontos szerepet játszik a nukleinsavak metilációja. Ennek tanulmányozására indították el a **Human Epigenome Projectet** (8,9). Ebből egy új tudományág nőtt ki, az **epigenomika**, amely a genom, illetve a genom mellett található fehérjék olyan módosulataival foglalkozik (pl. metiláció, acetiláció, hiszton módosulások), amely nem érintik közvetlenül a DNS szekvenciát, a bázissorrendet, de működésében fontos szerepet játszanak, sőt tovább is örökíthetőek. A metilációs mintázat hibái is vezethetnek betegségekhez. Ide tartoznak az imprintált géneknél említett szindrómák (az imprintáltság is a metiláció révén szabályozott), vagy pl. egyes tumorok, fragilis X vagy a Rett szindróma.

A humán és más élőlényének genomjáról számos web oldalon gyűjthetünk információkat, pl.: http://genome.ucsc.edu/; http://www.ensembl.org/; http://www.ncbi.nlm.nih.gov/.
A genetikai variációkkal, témánkban betöltött jelentős szerepük miatt a következő alfejezetben foglalkozom.
A humán genom feltérképezése a HGP hivatalos lezárása után sem fejeződött be. Megalakult a **Genome Reference Consortium**, amelynek fő feladata, hogy a genomban még megtalálható hiányokat („**gap**"-ek) feltérképezze, illetve az esetleges hibákat korrigálja. A

gap-ek a genom nehezen megszekvenálható, főleg ismétlődéseket tartalmazó régióiban találhatók. Becslések szerint a HGP befejezésekor, 350 ilyen gap volt, a szerkezei variációk egy része ezekben a gap-ekben található. Ez a régió nem kicsi, a teljes genom kb. 5%-ának felel meg. A feladat nehézségére jellemző, hogy 6 évvel később, 2009-ben még csak 50 ilyen gap-et sikerült megszekvenálni.

1.6. A humán genom variációi

Az emberi genom megismerését célzó Human Genome Project (HGP) része volt a humán genom variációinak a vizsgálata is, ami témánk szempontjából olyan nagy jelentőségű, hogy lényegét külön ismertetem (2,3).

A genom leggyakoribb variációja az egy nukleotidot érintő polimorfizmus, angol rövidítéssel SNP (single nucleotide polymorphism). Általában SNP-ről beszélünk akkor, ha a variáció populációs gyakorisága meghaladja az 1%-ot. Az ennél kisebb gyakoriságban előforduló variációt általában mutációnak szoktuk nevezni, bár ez utóbbi kifejezést főleg akkor használjuk, ha a variációnak fenotípusosan is megjelenő, funkciót módosító hatása is van. Azonban az elmúlt időben, az SNP-k meghatározása óriásit változott (ld. módszerek), és SNP-nek neveznek általában mindenféle egy nukleotidot érintő variációt, azzal hogy hozzáteszik, hogy mekkora a ritka allél gyakorisága, azaz a **MAF**-ja (**minor allele frequency**). Ennek a pontos definícióját még nem adták meg, de általában gyakorinak mondják, ha a MAF > 5%, alacsonynak, ha a MAF 0,5-5% között van, és ritkának ha < 0,5%, néhány publikációban ez utóbbi 0,3%.

A legtöbb SNP az intronokban található, utánuk következnek az intragenikus régiók, és végül legritkább az SNP az exonokban. Általában, átlagosan minden 1000 nukleotid polimorf, viszont az összes SNP-nek csak a 0,12-0,17%-a változtat meg aminosav kódot, és ennek is csak 40-47%-a non-konzervatív. A többi SNP első ránézésre semleges, azonban a legújabb vizsgálatok rámutatnak arra, hogy fenotípusos megjelenés szempontjából (ide tartozik pl. a betegségekre való hajlam, környezeti tényezőkre, gyógyszerekre való reagálás is) nem meghatározó, hogy az illető SNP változtat-e meg aminosav kódot vagy nem. Sőt az utóbbi időkben felfedezett betegségekhez kapcsolható SNP-k túlnyomó többsége nem változtat meg aminosav kódot (4).

Már a HGP során is találtak nagyobb szekvencia variációkat a genomban, azonban ezek jelentőségét populációs szinten, az SNP-kel összehasonlítva elhanyagolhatónak mondták. Azonban 2006 végén, ahogy egyre javultak a genom vizsgálati módszerei, rájöttek, hogy a genomban rengeteg kisebb–nagyobb méretű kópia szám variáció fordul elő (10). Azaz vannak olyan 1.000-től akár több megabázis nagyságú nukleotid szekvenciák, amelyek, ha a genomokat összehasonlítjuk, különböző kópia számban fordulnak elő. Ezekből ugyan nincs olyan sok, mint az SNP-kből, azonban, mivel nagyobb genomterületeket érintenek, összességében két ember között nagyobb variációért felelnek, mint az SNP-k. Ezt a típusú variációt elnevezték **copy number variation**-nak azaz **CNV**-nek (10). Legtöbbször a genom szerkezeti variánsait általában a CNV-khez szokták sorolni. A 3. ábra mutatja be a lehetséges szerkezeti variánsokat.

Becslések szerint a teljes genom 12%-át érintik ezek a variációk, és eddig 2.900 gént (a gének 13%-a) találtak, amely érintett. Ez azt jelenti, hogy egyes emberek különbözhetnek abban, hogy egy génből hány kópia található meg bennük. Az esetek többségében ez semmilyen látható tünetet nem okoz, de már számos betegségben igazolták a CNV-k szerepét. Ilyen pl. a skizofrénia, a HIV/AIDS hajlam, Crohn betegség, vesebetegségek, Alzheimer kór vagy az obezitás. Az SNP mintájára, ahol a polimorfizmus szó arra utal, hogy a variáció gyakorisága nagyobb mint 1%, bevezették a **copy number polymorphism** (**CNP**) kifejezést is a gyakori CNV-kre.

A betegségek mellett pl. a transzplantációban is szerepet játszhatnak a CNV-k. Például, vannak populációs szinten is gyakori, egész géneket érintő deléciók. Ilyenkor, ha a beültetést kapó szervezetből hiányzik egy gén, és így az abból expresszálódó fehérje, akkor, ha a donor szervben megtalálható ez a fehérje, az akceptor szervezet immunválaszt adhat a beültetett szerv ellen az MHC egyezés ellenére, pl. csontvelői őssejt átültetésnél „graft versus host betegség" léphet fel (11).

A CNV-kel kapcsolatban még egy érdekes felfedezést tettek. Általánosan elfogadott, hogy az egypetéjű ikrek genetikailag teljesen egyformák. Ezt a dogmát változtathatja meg az a felfedezés, hogy különbséget találtak a CNV-k tekintetében egypetéjű ikrek között (12). Ez utóbbi azt is mutatja, hogy szomatikusan is keletkezhetnek, pl. a magzati fejlődés során.

A CNV-k kimutatása szempontjából fontos, hogy a CNV-k egy részénél található olyan SNP, amely kapcsoltan öröklődik, azaz ezeknek a CNV-knek a kimutatásához elégséges a sokkal egyszerűbben (ld. módszerek fejezet) kimutatható SNP-ket detektálni.

A CNV-ket is figyelembe véve egy ember két genomja (itt a két szülőtől kapott kromoszómakészlete**) átlagosan 0,5%-ban különbözik egymástól**, azaz a CNV-k körülbelül 4x akkora különbségért felelnek, mint az SNP-k. Ezt az elsőnek megszekvenált ember, Craig Venter genomjából állapították meg ((13) 2. táblázat). Azonban a különbség történelmileg régen elvállt embercsoportok között akár a 3%-ot is elérheti. Érdekesség, hogy ezek a különbségek egy része véletlenszerű mutációk révén alakult ki (és ún. **random drift,** azaz véletlen sodródás során halmozódhat fel lokálisan), másik részük kialakulásában viszont a **természetes szelekció** játszott szerepet. Ide tartoznak pl. a bőrszínt, vagy az immunválaszt befolyásoló (baktériumok, vírusok által formált) genetikai variációk, melyek egy részét már sikerült beazonosítani (ld. gén-környezet fejezet).

A modern ember-genom fejlődéséről alkotott elképzelésünk jelentős változásokon ment keresztül az elmúlt években. 2010 májusában Scante Pääbo és munkatársai először a Neandervölgyi ember, majd egy nemrégiben felfedezett ember-populáció a **Denisova-i (**magyaros írással **gyenyiszovai) ember** genomját szekvenálta meg (14,15). Itt kell megjegyezni, hogy jelenleg még nem eldöntött kérdés, hogy ezek az embertől külön fajként definiálhatók-e, vagy ugyanannak a fajnak egy alfajának? Korábban az elmélet az volt, hogy a modern ember egy csoportja kb. 50.000 évvel ezelőtt elhagyta Afrikát, és benépesítette a Földet. Azonban, ezekben a vizsgálatokban azt találták, hogy a modern ember a Közel-Keleten részben keveredett a **Neandervölgyi ember**rel, illetve bizonyos embercsoportok a Denisova-i emberrel. Ennek következtében bizonyos ma élő embercsoportok 1-4%-ban a Neandervölgyi, a Pápua Új Guinea-n és egyes szigeteken élők, ausztrál őslakosok, óceániaiak stb. pedig a Denisova-i ember genomjának 4-6%-át hordozzák. Pl., a melanézek, mindkét populációtól hordoznak genom-nyomokat, genomjuk kb. 8%-ában. Ezek nyilván hozzájárulnak két ember genomja közötti különbségekhez (16,17).

Érdekességként itt lehet megjegyezni, hogy egyes kutatások alapján ez a hozzákeveredés, vagy angolul admixture, pozitív hatással volt a ma élő ember immunrendszerére. Becslések szerint a ma élő ember HLA alléljainak kb. 50% az archaikus emberektől származik, növelve ezzel a patogének felismerésében fontos szerepet betöltő HLA variációinak számát, így populációs szinten a faj stabilitását (ld. gén-környezet kölcsönhatás fejezet).

A HGP után a HapMap projektek és az **1000 genom projekt** járultak hozzá jelentős mértékben az emberi genom variációinak feltérképezéséhez. Például a hét populációt vizsgáló 1000 genom projekt első eredményeit bemutató „pilot" cikkében 15 millió SNP-t, 1 millió rövid inzerciót, vagy deléciót és 20 ezer szerkezeti variánst közöltek. Ezek többsége új variáció volt. Becslések szerint azonosították az ezekben a populációkban található összes variáció 95%-át (8).

A variációk és a gének száma szempontjából kiemelkedik az emberi genom 6p21.3 régiójában található MHC (vagy HLA) régió. Ezen a 7,6 Mb szakaszon kódolódnak az

immunválaszban, transzplantációban, véradásnál létfontosságú szerepet betöltő MHC, vagy HLA gének. Ennek a genomterületnek az ún. class III régiójában található a legnagyobb génsűrűség (58 expresszálódó gén), illetve az egész régióra jellemző a nagyfokú diverzitás. Egy vizsgálatban ennek egy 4 Mb-nyi régiójában 37 ezer SNP-t és 7 ezer szerkezeti variánst találtak, ami kb. egy nagyságrenddel nagyobb variáció sűrűség, mint a genom többi részén.

1.7. „Junk DNS" a humán genomban

Az alacsony génszám még a szakembereket is meglepte, sőt szinte sokként érte. Erre jellemező, hogy még 2000-ben a Cold Spring Harbor Laboratory Genome Meeting-en a terület specialistái fogadtak a humán genom génszámára. A számok 26 ezer és több mint 150 ezer között terjedtek, így végül a legalacsonyabb számot tippelő nyerte a fogadást, pedig még ő is kb. 20%-kal magasabb számot tippelt a valóságosnál. Az eredmény azért is meglepte a szakembereket, mert például az alig 1 mm nagyságú, szabad szemmel gyakorlatilag láthatatlan, igen egyszerű felépítésű Caenorhabditis elegans nevű fonálféreg, amely az egyik legnépszerűbb modell állat biológiai kísérletekben, nagyságrendileg hasonló számú gént tartalmaz. Ebből az alacsony génszámból következik, hogy a humán genom fehérjét kódoló részének aránya mindössze **1,2%-a a teljes genomnak**. Mivel korábban úgy gondolták, hogy a genom fő feladata az, hogy fehérjéket kódoljon, a fehérjét nem kódoló részt a genom hulladékának, angolul „junk"-nak nevezték (18). A szakemberek azonban nyilvánvalóan érezték, hogy az emberi genom egyszerűen nem állhat 98,8%-ban szemétből, felesleges szekvenciából! Hogy tisztázzák ezt az ellentmondást, 2003-ban elindították a Encyclopedia of DNA elements (**ENCODE) projekt**et: http://genome.ucsc.edu/ENCODE/ , http://www.genome.gov/10005107, amely az első fázisban azt tűzte ki maga elé, hogy a genom 1%-ban felderíti az összes funkcionális egységet (19). Ez a projekt 2007-ben lezárult, de 2008-ban 80 millió $-os költségvetéssel további 4 évre meghosszabbították. Az eredmények lényege, hogy az emberi genom kb. 5%-a evolúciósan konzervált, azaz tőlünk igen távoli fajokkal szinte tökéletesen megegyezik, azaz a nem-fehérjekódoló régióknak is nyilvánvalóan van valamilyen funkciója (pl. transzkripciós faktor kötőhely, regulátor szekvencia, miRNS kötőhely, RNS-t kódol stb.), illetve a genom 3 dimenziós szerkezetének és kromatin struktúrájának is fontos jelentősége van. Ez utóbbira is történtek vizsgálatok. Eszerint, ha olyan algoritmussal hasonlították össze a különböző fajok genomját, hogy az azokat alkotó nukleinsavak milyen 3D szerkezetet vesznek fel, akkor a humán genom 12%-át találták konzerváltnak (20). Egyes elméletek szerint, építőipari hasonlatot használva a genomban kódolt fehérjéket nevezhetjük építőköveknek, míg az azon kívüli részek tartalmazzák azt az információt, hogy ezeket a köveket hogyan kell úgy összerakni, hogy azokból egy működőképes szervezet jöjjön létre. Erre egyfajta bizonyíték az is, hogy bár az élet már 4 milliárd évvel ezelőtt kialakult a Földön, a többsejtű élőlények mindössze 525 millió évvel ezelőtt jelentek meg. Valószínűnek tűnik, hogy ez a 3,5 milliárd éves evolúció kellett ahhoz, hogy egy olyan szabályozó mechanizmus alakuljon ki, amely lehetővé tette a bonyolultabb életformák létrejöttét. Nyilvánvaló, hogy a szabályozáshoz szükséges információ nem a fehérjéket kódoló génekben található, azaz elvileg minél több ilyen szekvencia van egy genomban, annál több szabályozó mechanizmus „férhet bele" (18).

Az alacsony génszám-bonyolult szervezet ellentmondást némileg feloldja az a felfedezés is, hogy génjeink 94%-a nemcsak egy fehérjét kódol, azaz például alternatív splicing útján a különböző szövetekben ugyanabból a génből más-más szerkezetű és funkciójú fehérjék íródnak át. Egyszerűbb szervezeteknél ilyen mechanizmus nincs, vagy csak jóval kisebb mértékű. Továbbá, a fehérjék poszt-transzlációs módosításaival rengeteg különböző fehérje jöhet létre. Egyes becslések szerint az ember fehérjéinek száma eléri a 2 milliót.

2001-ben még úgy gondolták, hogy a genom fő funkcionális részei a fehérjéket kódoló gének. Az **RNS-ek**et egyfajta segédszereplőknek gondolták, azaz fő funkciójuk, hogy (mRNS, t-RNS, rRNS formában) a fehérjére való átíródásban segédkezzenek. Azonban az utóbbi években egyre másra fedezik fel, hogy az RNS-eknek milyen funkcióik vannak még, főleg a szabályozásban. A leghíresebb példa az RNS interferencia és a mikro RNS-ek (**miRNS**) felfedezése, amelyért 2006-ban Nobel-díjat is adtak (21). Kiderült, hogy a gének > 60%-ának a szabályozásában játszanak szerepet, egy miRNS akár több száz génében is, lehetővé téve egy igen bonyolult, összehangolt szabályozást. De ide tartoznak a főleg a spermatogenezis szabályozásában szerepet játszó *piwi-interacting* (pi) RNS-ek, vagy a pszeudégenről átíródó *competitive endogenous* RNS-ek (ceRNS), vagy *antisense termini-associated short* RNS-ek (aTASRs), *Large intervening noncoding* RNS (lincRNS), stb. (22). Általában ezek fő szerepe a transzkripción, ritkábban a transzláción keresztüli géncsendesítés. Korábban szintén nem gondolták, hogy a teljes genom 74-93%-a átíródik RNS-sé, és valószínű, hogy ez nem történik véletlenül.

1.8. Komparatív genomika

A junk-nak nevezett, fehérjére nem átíródó szekvenciák fontos szerepét mutatják a komparatív genomika eredményei. A komparatív genomikában különböző fajok genomját hasonlítják össze, és olyan kérdésekre keresik a választ, mint pl.: Milyen gének jellemzőek egy fajra (pl. csimpánz vs. ember)? Milyen szekvenciák nélkülözhetetlenek az emlős élethez (pl. egér vs. ember vs. gyümölcslégy)? Melyek a többsejtűek alapfehérjéi (féreg vs. ember vs. egysejtűek)? stb. Ezek a mi témánk szempontjából olyan eredményeket hoztak, mint pl., hogy az egér hasonlósága az emberrel 90%, génjeinek 99%-ának van humán megfelelője, csak kb. 300 génben különbözünk, azaz az egér jól használható mint modell élőlény emberi gének funkciójának, vagy betegségek vizsgálatában (23).
Az egér az evolúciós fejlődésben kb. 75 millió évvel (420 millió egérgeneráció) vált el tőlünk. Érdekes módon, a kutyától régebben váltunk el, azonban, tekintve a kutya lényegesen hosszabb generációs idejét, gén-szinten kisebb a különbség. Az emberi genomot ezzel a két állatfajjal összehasonlítva 7-7,5%-ban találtak konzervált régiókat a fehérjéket kódoló géneken kívül (24).
Konzervált régiók: Ha két genomikai régió egymással csaknem teljesen megegyezik két, egymástól már régen elvált fajban, akkor ez a régió szelekciós nyomás alatt van, azaz valamilyen olyan funkciója van, amelynek változása életképtelenné teszi az élőlényt (a mutáns kiszelektálódik).
Nyilvánvalóan az ember legközelebbi rokonánál, a csimpánznál, melytől >6.3 millió évvel, és mindössze kb. 250 ezer embergenerációval ezelőtt váltunk el még több azonosságot találtak. Ennek közelségét egy olyan, kicsit humoros példával szokták demonstrálni, hogy képzeljük el, hogy egy ember ha megfogja az anyja kezét, és az is az ő anyjáét, és így tovább, akkor ha a mai ember Budapesten van, akkor kb. Debrecenben már egy csimpánz-szerű anya fog állni. A csimpánzzal hasonlóságunk 99%, ha az egymáshoz illeszthető szekvenciákat nézzük, összességében a hasonlóság 96%. A különbségekért főleg inzerciók/deléciók („indelek") a felelősek. Érdekes módon a legnagyobb a különbség az Y kromoszómában, ahol a két faj 30%-ban különbözik. Az X kromoszóma hasonlósága miatt viszont, egy akkoriban nagy port felvert feltételezés is történt, azaz 1,2 millió évvel a két faj szétválása után kimutatható még „genomcsere", azaz utódokat eredményező párosodás történt a két faj között. Egy másik érdekesség, hogy találtak egy, amúgy nagyon konzervált gént (**FOXP2**), amely különbözött az ember és a csimpánz között, de nem az ember és a neandervölgyi ember (*Homo neanderthalensis*) között. A FOXP2 mutációja emberben egy sajátos beszédzavart

okoz, azaz a mutáció hordozója képtelen a nyelvtan legalapvetőbb szabályait is megtanulni. Ezt annak idején elnevezték (nyilván helytelenül) nyelvtan génnek.
Ha a ma élő ember genomját a régen kihalt neandervölgyi ember genomjával hasonlították össze, aminosav szekvenciában mindössze 1000-2000 különbséget találtak, a különbség a csimpánzokhoz képest 20-50x kevesebb. Igaz, ahogy előzőekben már tárgyaltuk, a korábbi tudományos vélekedéssel ellentétben a két faj szétválása után (amely 500 ezer évvel ezelőtt történt) párosodott egymással, azaz mind az európaiak, mind az ázsiaiak genomjának 1-4%-n kimutatható a két faj keveredése. 78 olyan fehérjeváltozást okozó különbséget találtak az emberi genomban, amely ez idő alatt alakult ki, illetve jó néhány olyan változást, amely az emberben pozitív szelekciónak minősülhet. Ilyenek pl. a sperma mozgékonyságot, a sebgyógyulást, a bőr működését, vagy a kognitív képességeket befolyásoló („javító”) mutációk (17, 25).

Legutolsó frissítés	2011. szeptember
Genom nagysága (bázispár)	3.283.984.159
Ismert fehérjét kódoló gén (darab)	20.469
Pszeudogén (darab)	14.266
RNS gén (darab)	12.499
Gén exon (darab)	640.185
Rövid variánsok, pl. SNP (darab)	30.099.223
Szerkezeti variáns	1.772.315

1. táblázat

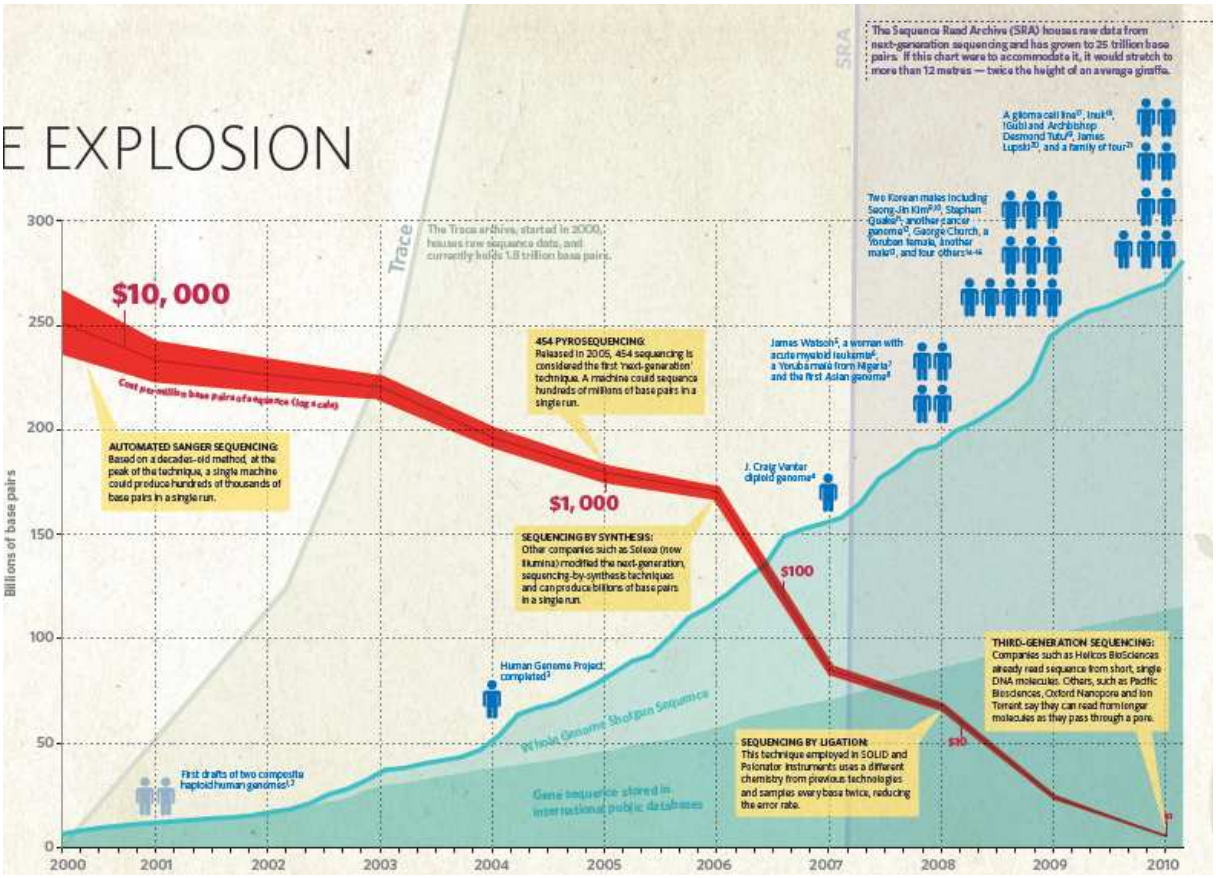
Néhány statisztikai adat a humán genomról. Forrás: http://www.ensembl.org/Homo_sapiens/Info/StatsTable

	SNP-k száma	
J. Craig Venter genomja	3,213,401	
James Watson genomja	3,322,093	
Ázsiai genom	3,074,097	
Yoruban (afrikai) genom	4,139,196	
	Szerkezeti variánsok Venter genomjában	
	Darab	**Hossz (bp)**
CNV	62	8.855–1.925.949
Inzerció/deléció	851.575	1–82.711
Blokk szubsztitúció	53.823	2–206
Inverzió	90	7–670.345

2. táblázat

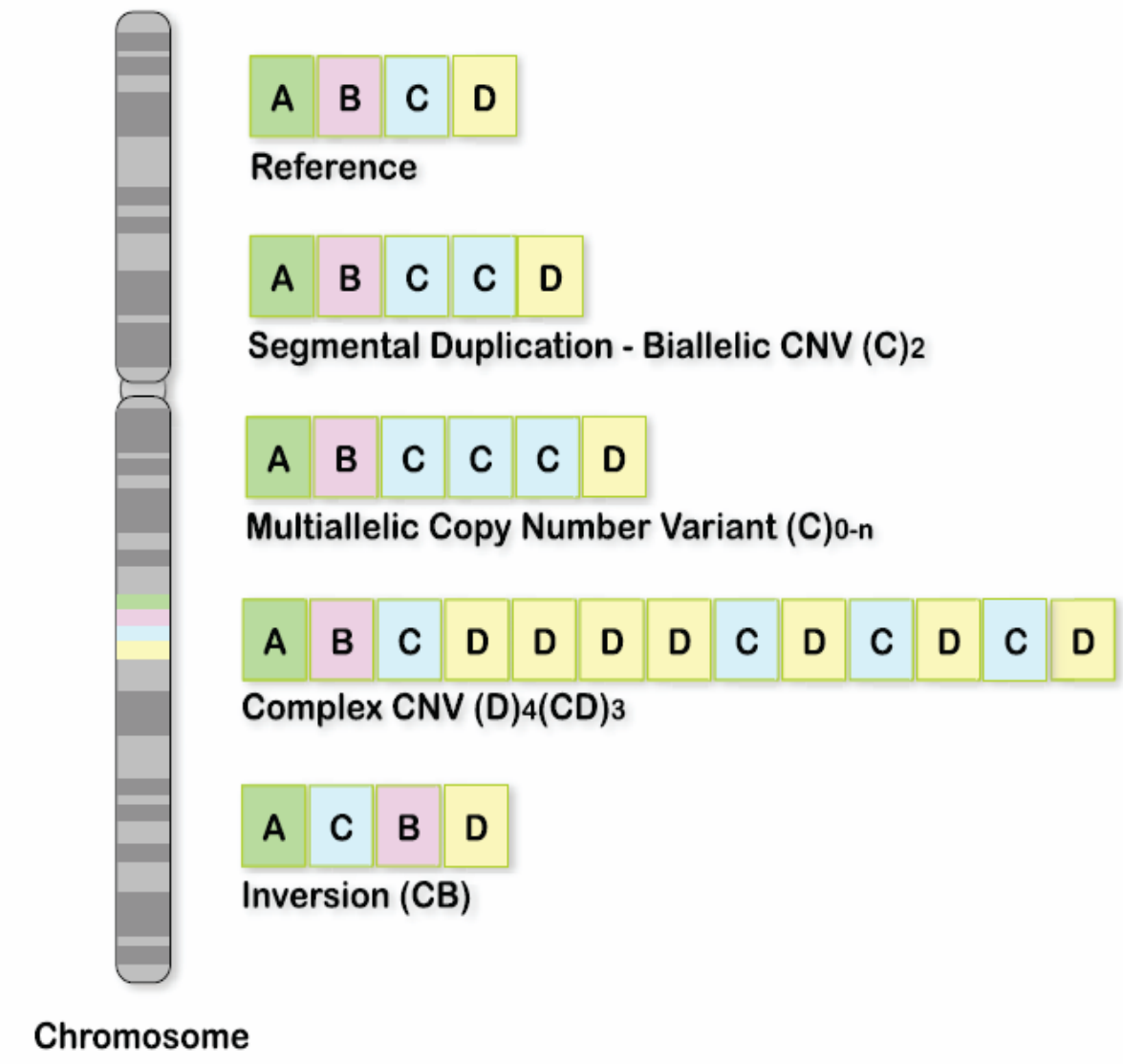
Genetikai variációk aránya különböző megszekvenált humán genomokban (13).

1. ábra
A genomikai X díjnak az emblémája (5)



2. ábra

DNS szekvenálás árának (piros vonal) és az adatbázisokban tárolt befejezett DNS szekvenciák mennyiségének (kék vonal) változása 2000 és 2010 között logaritmikus skálán. 2000 környékén egy millió bázispár megszekvenálása 10 ezer $-ba került, amely 2010-re már 1 $-ra csökkent. A befejezett DNS szekvencia mennyisége 2000-ben 8 milliárd bázispárral indult, és kb. minden 18. hónapban megduplázódott a mennyisége. 2010-re 270 milliárd bázispárra nőtt. Ez azonban eltörpül a nyers adatok mennyisége mellett, amelyet a *Trace archive and Sequence Read Archive* (SRA)-ban tárolnak. Az itt tárolt mennyiség 25 trillió bázispár volt 2010-ben, amit, ha ebben a koordináta rendszerben akarnánk ábrázolni, 12 méterre lógna ki a könyvből, ami kétszerese egy átlag zsiráf magasságának. A Nature 2010:464:671 cikk ábrája alapján.



3. ábra
A genomban előforduló szerkezeti variánsok

1.9. 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 et al. 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. Rusk N, Kiermer V.Primer: Sequencing—the next generation. Nature Methods 2008;5:15.
7. http://www.genome.gov/10005107; 2009.
8. Pennisi E. 1000 Genomes Project Gives New Map Of Genetic Diversity. Science 2010; 330: 574-5.)
9. http://www.epigenome.org/; 2009.
10. Redon R. és mtsai.:Global variation in copy number in the human genome. Nature 2006; 444: 444-454.
11. Armour JA. Copy number variation and antigenic repertoire. Nat Genet. 2009;41(12):1263-4.
12. 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.
13. Ng PC, et al. Genetic variation in an individual human exome. PLoS Genet. 2008 Aug 15;4(8):e1000160.
14. Reich D, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010 Dec 23;468(7327):1053-60.
15. Green RE, et al. A draft sequence of the Neandertal genome. Science. 2010 May 7;328(5979):710-22.
16. Reich D, et al.Denisova admixture and the first modern human dispersals into southeast Asia and oceania. Am J Hum Genet. 2011 Oct 7;89(4):516-28.
17. Burbano HA, et al. Targeted investigation of the Neandertal genome by array-based sequence capture. Science. 2010 May 7;328(5979):723-5.
18. Gibbs W.W. (2003) "The unseen genome: gems among the junk", Scientific American, 289(5): 46-53.
19. 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.
20. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. Local DNA Topography Correlates with Functional Noncoding Regions of the Human Genome. Science. 2009; 324: 389 – 392.
21. Fire A, Xu S, Montgomery M, Kostas S, Driver S, Mello C (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans". Nature 391 (6669): 806–11.
22. Swami M. RNA world: A new class of small RNAs Nature Reviews Genetics 2009;10, 425.
23. Waterston RH. Et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002; 420 (6915) 520 - 562.
24. Kirkness EF et al. The Dog Genome: Survey Sequencing and Comparative Analysis. Science. 2003; 301:1898-1903
25. Krause J et al. The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals. Current Biology 2007; 17: 1908-1912

1.10. Fejezethez tartozó kérdések

1. Mi az a genom?
2. Mivel foglalkozik a genomika?
3. Mi a különbség a genetika és a genomika között?
4. Mikor indult a Humán Genom Projekt?
5. Melyik szervezet koordinálta a nemzetközi Humán Genom Projektet?
6. Mondjon néhányat a Humán Genom Projekt fő céljai közül!
7. Mi annak a cégnek a neve, amely 1998-ban kezdte meg a humán genom szekvenálását?
8. Miért adják az Archon X díjat a genomikáért?
9. Mondjon példát genom projektekre!
10. Mekkora kb. a humán genom mérete?
11. Hány fehérjét kódoló gént tartalmaz körülbelül a humán genom?
12. Hány százaléka körülbelül a fehérjekódoló rész a teljes genomnak?
13. Melyik a legnagyobb humán gén?
14. Melyik fehérjének van a leghosszabb kódoló szekvenciája?
15. Melyik kromoszómán legnagyobb a génsűrűség?
16. Melyik a leggénszegényebb kromoszóma?
17. Mikor keletkezik a legtöbb öröklődő mutáció?
18. Mit jelent az 1 cM?
19. Mit nevezünk CpG szigeteknek?
20. Mi az a genetikai imprinting?
21. Hogy hívják a leggyakoribb ismétlődő szekvenciát?
22. A humán genom kb. hány százaléka tartalmaz ismétlődő szekvenciát?
23. Egy ember átlagosan hány olyan gént hordoz, melyben funkcióvesztéses mutáció van?
24. Mi az az SNP?
25. Minek a rövidítése a MAF?
26. Mi az a Copy Number Variations (CNV)?
27. Mi az a CNP?
28. Átlagosan mennyire különbözik két ember egymástól genetikai szinten?
29. Keveredett-e a homo sapiens más emberfajokkal/alfajokkal?
30. Melyik a humán genom legvariábilisabb régiója?
31. Mit nevezünk pszeudogénnek és lehet-e funkcionális szerepe?
32. Hogyan keletkeznek a paralógok?
33. Mivel foglalkozik az epigenomika?
34. Hol gyakoribb a polimorfizmus az intronban vagy az exonban?
35. Mi az a junk DNS és mi a jelentősége?
36. Milyen szerepei lehetnek az RNS-eknek?
37. Mivel foglalkozik a komparatív genomika?
38. Mik azok a konzervált genomrégiók, és mi a jelentőségük?