



37. OTDK Biológia Szekció

2025. április 15-17.

**A hőindukált GUT15/CR20 lncRNS család vizsgálata
bioinformatikai módszerekkel**

Bioinformatic analysis of the heat induced GUT15/CR20 lncRNA family

Készítette: Gál Nóra

JKUW77

Biológia Doktori Iskola

nappali tagozat

Szent István Campus

Témavezető: Dr. Csorba Tibor

Tudományos főmunkatárs, csoportvezető

Genetika és Biotechnológia Intézet, Növénybiotechnológia tanszék,

Növényi Stresszbiológia csoport

Gödöllő

2025.01.08.

Tartalomjegyzék

1. Bevezetés	4
2. Célkitűzés	5
3. Irodalmi áttekintés	5
3.1. Az lncRNS-ek	5
3.1.1. GUT15/CR20 lncRNS család	8
3.2 lncRNS-ek vizsgálati lehetőségei	9
3.2.1. Szekvenciaillesztés	10
3.2.2. Filogenetikai elemzés	10
3.2.3. K-mer analízis	11
4. Anyag és módszer	12
4.1 Számítógépes környezet	12
4.2. Szekvenciaillesztés	12
4.3. Filogenetikai elemzés	13
4.4. Másodlagos szerkezet meghatározása	13
4.5. K-mer analízis	14
4.6. Hierarchikus klaszterelemzés	14
4.7. ORF-ek azonosítása	14
5. Eredmények és megvitatásuk	16
5.1. Szekvenciaillesztés és filogenetikai elemzés	16
5.1.1 Szekvenciaillesztés	16
5.1.2 Filogenetikai fa készítése	20
5.1.3. A HCR régió szerkezeti elemzése	23
5.2. K-mer analízis, hierarchikus klaszterelemzés	24
5.2.1. A HCR régió K-mer tartalmának vizsgálata	25
5.2.2. <i>Triticum</i> fajok K-mer analízise és hierarchikus klaszterelemzése	26
5.2.3. A HCR régió kívüli szakaszok K-mer tartalmának vizsgálata	30
5.2.4. A cDNS/ncRNS szekvenciák K-mer analízise, hierarchikus klaszterelemzése és a kialakult csoportok vizsgálata	32
5.2.5. <i>Arabidopsis thaliana</i> ncRNS szekvenciák K-mer analízise, hierarchikus klaszterelemzése és a kialakult csoportok vizsgálata	33
5.3 sORF-ek azonosítása a cDNS és ncRNS szekvenciákban, sORF funkcionalitás lehetőségének tesztelése	34
6. Összefoglalás	35

7. Köszönetnyilvánítás	36
8. Irodalomjegyzék.....	37

1. Bevezetés

A magasabb-rendű eukarióták genomjának túlnyomó része átíródik RNS-sé, azonban az összes RNS (tehát a teljes transzkriptom) csak töredéke kódol fehérjét. A transzkriptom azon része, ami nem fehérjét kódol RNS-ként funkcionálhat. Az 1950-es években és ezt követően, számos fehérjét nem kódoló RNS (ncRNS) családot fedeztek fel; ilyenek a transzfer RNS-ek (tRNS), riboszomális RNS-ek (rRNS), kis magi RNS-ek (small nuclear RNA, snRNS), kis RNS-ek (siRNA) és a hosszú nem-kódoló RNS-ek (long non-coding RNA, lncRNS). Míg a tRNS, rRNS, snRNS, sRNS stb. családok funkciója jól ismert, az lncRNS-ek funkcióiról továbbra is hiányosak az ismereteink.

A korai feltételezések szerint az lncRNS-ek transzkripcionális melléktermékek (“transcriptional noise”) lehetnek. Egyrészt mivel sok közülük nagyon alacsony szinten fejeződik ki, másrészt pedig jellemzően nem konzerválódtak a törzsfejlődés során, sok esetben még a közeli fajokban sem, éppen ezért az evolúciós kiválogatódás nem alakíthatott ki funkcionalitást bennük. Azonban számos lncRNS esetén megállapították, hogy jelenléte/abundanciája szövet-, szerv-specifikus, mennyisége pedig változhat belső (pl. fejlődési állapot, hormonhatások), vagy külső környezeti tényezők (pl. biotikus vagy abiotikus stresszhatások) hatására. Emellett sok lncRNS érési folyamatokon (pl. intron kivágódás) is keresztül megy. Ezek a megfigyelések arra engednek következtetni, hogy az lncRNS-ek szabályozott módon fejeződnek ki, következésképpen biológiai relevanciájuk lehet.

Bár ezidáig több ezer lncRNS-t azonosítottak különféle organizmusokban, konkrét szerepet csak néhány tucatnyi esetében sikerült bizonyítani. Elképzelhető, hogy az lncRNS-ek egy aktívan evolválódó nukleinsav csoport, melyből az evolúció folyamatosan “merít” és ruház fel szabályozó funkcióval. Ebben a szemléletben tehát az lncRNS-ek között akadhat non-funkcionális (transzkripció melléktermék), kialakulóban lévő “*szeml*”-funkcionális (gyenge aktivitást hordozó) vagy funkcionális molekula.

Munkánk során egy, a zárvatermőkben jelenlévő lncRNS családot vizsgáltunk. Ezekben az lncRNS molekulákban egy rövid szakasz szekvenciája konzervált (kb. 120 nukleotid (nt)), a fennmaradó rész (1000-2500 nt) azonban magas variabilitású nem-konzervált régió. Kísérleteinkben bioinformatikai módszerekkel elemeztük a konzervált régió jelenlétét, szerkezetét és evolúciós eredetét, illetve a nem-konzervált régiók mikrokonzerváltságát, variabilitását és filogenetikai jellemzőit. A munka folyamán kialakított bioinformatikai eszköztár használata irányadó lehet a

valós biológiai funkció teszteléséhez és megállapításához, a vizsgált lncRNS család evolúciós folyamatainak jobb megértéséhez, de segítséget nyújthat általánosan, más lncRNS-ek hasonló vizsgálatában is.

2. Célkitűzés

Korábbi munkánk során molekuláris biológiai és genetikai módszerek segítségével azonosítottunk egy hőre változó konzervált lncRNS családot *Arabidopsis thaliana* modell-növényben. A család három tagja a *GUT15* (*Gene with Unstable Transcript 15*), a *CR20* (*Cytokinin-repressed 20*), illetve egy részlegesen konzerválódott homológ lókuszt (*Atlg19397*, egy pszeudogén, mely nem fejeződik ki). A két előbbi családtag (GUT15, CR20) magas hőmérséklet hatására erőteljesen halmozódik, ami arra enged következtetni, hogy szerepük lehet a hőmérsékleti változások érzékelésében és/vagy a hőstressz válasz szabályozásában.

Munkánk célja, hogy a korábbiaktól eltérő, *bioinformatikai* módszerekkel vizsgáljuk a GUT15/CR20 lncRNS családot:

- (i) meghatározzuk a GUT15/CR20 lehetséges homológjait az egyes növényfajokban és megvizsgáljuk a homológok közötti hasonlóságot és eltéréseket;
- (ii) növényekben és más eukariótákban ismert lncRNS tulajdonságokkal való összevetéssel, kísérletet tegyünk a (fehérje)partner kötőhelyek (RNS domének, *cis*z elemek) azonosítására és ezáltal a feltételezhető funkció meghatározására;
- (iii) az lncRNS homológjainak azonosításával és összehasonlításával feltérképezzük a lehetséges evolúciós utat, mely a GUT15/CR20 lncRNS család kialakulásához vezethetett (2. ábra);
- (iv) mindezekon túlmenően, egy olyan automatizált bioinformatikai munkafolyamatot (“work-flow”-t) szeretnénk létrehozni, amely alkalmazásával átláthatóbb és könnyebben vizsgálható más lncRNS-ek konzerváltsága, funkcionalitása és evolúciója is.

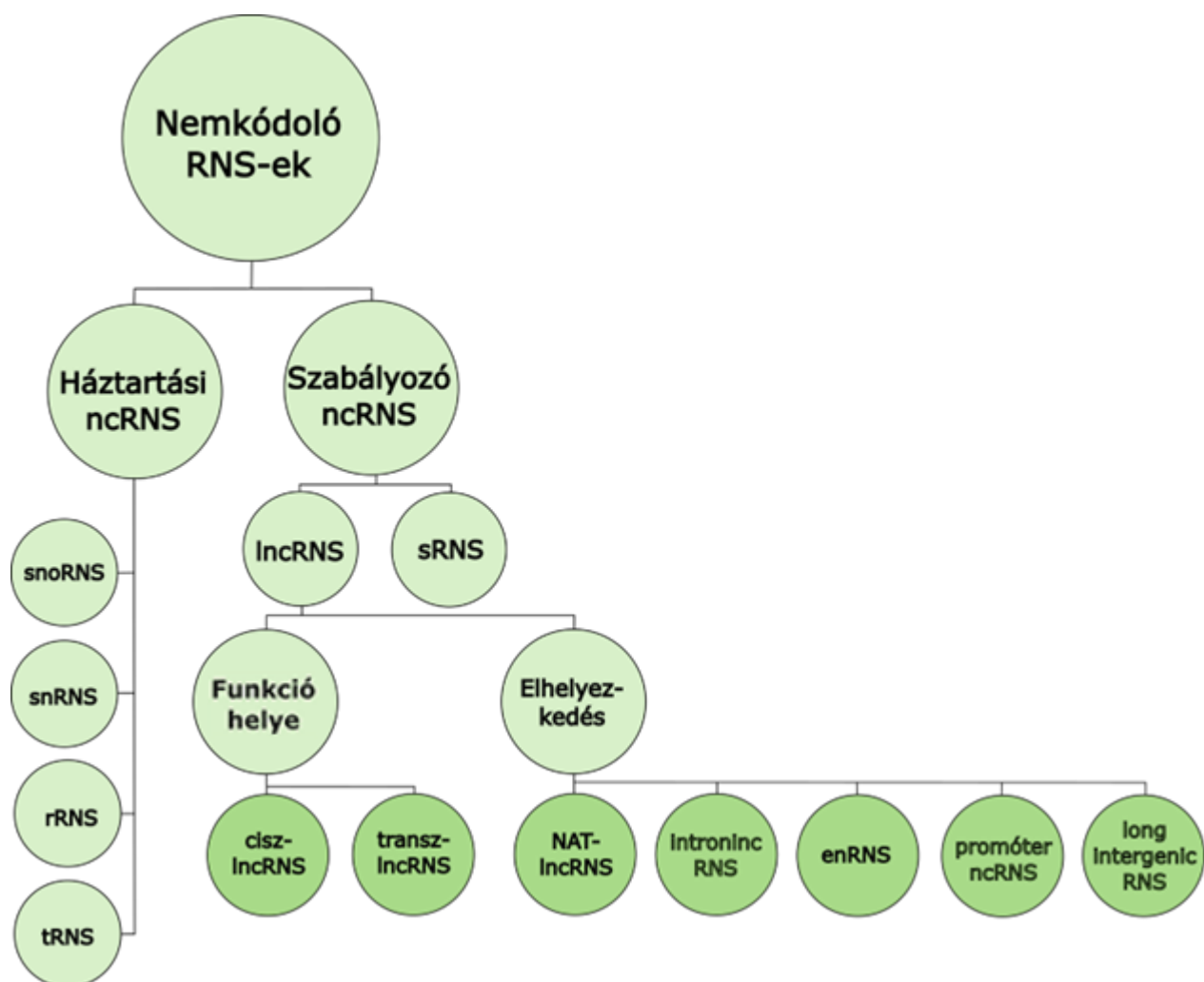
3. Irodalmi áttekintés

3.1. Az lncRNS-ek

A **nem-kódoló RNS-ek (ncRNS-ek)** kutatása a bioinformatika és az újgenerációs szekvenálási módszerek fejlődésének köszönhetően egyre nagyobb teret kapott az elmúlt években (Wang és

mtsai., 2017). Vizsgálatuk jelentőségét igazolja, hogy a magasabbrendű eukarióta genomok akár 70-90%-a RNS-re íródhat át, azonban ezeknek csupán töredékéről (pl. humán genom 2%-áról, ami kb. 20-25,000 gén) készül fehérje termék (Rai és mtsai., 2019). A fennmaradó RNS átiratok mind szerkezet, mind funkció tekintetében rendkívül változatos molekulák, emiatt csoportosításuk nem minden esetben egyértelmű (Cech & Steitz, 2014; Mattick és mtsai., 2023; Peschansky & Wahlestedt, 2014). Alapvetően két nagy csoportjukat különböztetjük meg, a háztartási, strukturális funkcióval rendelkező– és a szabályozó ncRNS-eket. A háztartási RNS-ek legismertebb képviselői a transzfer RNS-ek (tRNS) és a riboszomális RNS-ek (rRNS), de ide tartoznak a kis magi RNS-ek (snRNS) és a kis nukleoláris RNS-ek (snoRNS) is. A szabályozó funkciót ellátó ncRNS-ek közé tartoznak a kis RNS-ek (sRNS, 21-24 nukleotid (nt) hosszúságúak)(Kruszka és mtsai., 2012; Sunkar és mtsai., 2012) és a hosszú nem-kódoló RNS-ek (lncRNS)(1. ábra)(Mattick és mtsai., 2023; Morris & Mattick, 2014).

Az ncRNS-ek csoportján belül a **hosszú nem-kódoló RNS-ek (lncRNS)** egy tetszőleges csoportot alkotnak (konvencionálisan a >200 nt hosszúságúak), ebből adódóan heterogén funkcióik vannak. Alacsony fejlettségű szinten lévő organizmusok néhány lncRNS gént hordoznak, míg a magasabbrendűekben számuk exponenciálisan nő (pl. a humán genomban akár 100,000 lncRNS gént feltételeznek). Az lncRNS-ek többsége alacsonyan konzervált, funkciójuk sok esetben egyedi. Az lncRNS-eket öt csoportba oszthatjuk a genomi elhelyezkedésük alapján; természetes antiszensz lncRNS-ek (natural antisense transcripts, NAT-lncRNAs), intron régióról származók (intronic ncRNA), enhancer régióról származó (enRNAs), promoter-asszociált transzkriptumok (promoter ncRNAs) és hosszú intergénikus ncRNS-ek (long intergenic RNAs, lincRNAs)(Shih & Kung, 2017). Az lncRNS-ek hatása kifejeződik a keletkezési helyükön, vagyis “*cis*”-ben, vagy távoli lokuszokon azaz “*transz*”-ban. A *cis*-lncRNS-ek esetében a hatás következhet magából a molekula keletkezésének folyamatából (transzkripció vagy RNS-érés folyamata) vagy az RNS molekula által, míg a *transz*-lncRNS-ek esetén önmaguk, vagy speciális esetben a róluk keletkező “mikro-peptid” a funkcionális egység (Choi és ., 2019). Az lncRNS-ek általában nem önmagukban, hanem több molekulából összeszerelődő komplexek alegységeiként működnek, partnereik lehetnek specifikus genomi lokuszok (DNS), RNS-ek vagy fehérjék. Ezek a komplexek epigenetikai, transzkripcionális, poszt-transzkripcionális, transzlációs szinten szabályozhatnak számtalan folyamatot.



1. ábra: A nemkódoló RNS-ek csoportosítása

Az ncRNS-ek fehérjét nem kódoló átíratok. Ezek lehetnek strukturális (háztartási), vagy szabályozó funkcióval rendelkező RNS-ek is. A háztartási ncRNS-ek közé tartoznak a transzfer RNS-ek (tRNS), riboszomális RNS-ek (rRNS), a kis-magi RNS-ek (snRNS) és a kis nukleoláris RNS-ek (snoRNS) is. Ezek aktívan részt vesznek az alapvető sejtfunciók ellátásában. A szabályozó ncRNS-ek pontos funkciója a legtöbb esetben (még) kérdéses, csoportosításuk emiatt méretük alapján történik. Megkülönböztetünk kis RNS-eket (sRNS), amik mérete 21-24 nt közé esik, és hosszú nem-kódoló RNS-eket (lncRNS), amik jellemzően 200 nt-nál hosszabbak. Az lncRNS-ek osztályozása történhet elhelyezkedésük (érési helyük) alapján, így megkülönböztetünk természetes antiszensz (NAT lncRNS), az intron régióról származó (intronc RNS), az enhancer régióról származó (enRNS), promóter asszociált (promóter ncRNS) és hosszú intergénikus RNS-eket (long intergenic RNS). Funkcionálásuk helye alapján pedig lehetnek cisz-lncRNS-ek, amik hatása helyben érvényesül és lehetnek transz-lncRNS-ek, amik a kifejeződésükhöz képest máshol fejtenek ki szabályozó hatást.

Mára több ezer lncRNS-t azonosítottak a **növényi organizmusokban** (*Arabidopsis thaliana*, repce, paradicsom, árpa, búza stb.) is. Legtöbb funkciója ismeretlen, azonban néhány lncRNS-t alaposan jellemeztek, főleg az *A. thaliana* modell szervezetben. Az APOLO például fontos szerepet játszik az auxin jelátvitelben és a gyökérfejlődés szabályozásában (Ariel és mtsai., 2014, 2020; Moison és

mtsai., 2021), az asDOG1 a DOG1 gént antiszenszként szabályozza, és csírázás szabályozásában vesz részt (Fedak és mtsai., 2016; Kowalczyk és mtsai., 2017). További példa még a FLORE, amely a cirkadián ritmus szabályozásában vesz részt (Henriques és mtsai., 2017), valamint a COLDAIR és COOLAIR, amelyek a FLC gén lokuszának szensz és antiszensz transzkriptjei, és fontos szerepet játszanak a virágzási idő és a vernalizáció szabályozásában (Henriques és mtsai., 2017; Heo & Sung, 2011; Swiezewski és mtsai., 2009), aminek fontos résztvevője emellett a FLAIL is (Jin és mtsai., 2023). Ismerünk a hőstressz válaszban részt vevő lncRNS-eket is, az asHSFB2a például az HSFB2a gént annak antiszenszeként szabályozza, és fontos szereplője hőstressz szabályozásnak (Wunderlich és mtsai., 2014).

3.1.1. GUT15/CR20 lncRNS család

A záratermőkben megtalálható GUT15 (*Gene with unstable transcript 15*) és CR20 (*Cytocinin-repressed 20*) lncRNS-eket közel 30 évvel ezelőtt azonosították dohány, illetve uborka fajokban (Taylor & Green, 1995; Teramoto és mtsai., 1995).

Taylor és Green rövid féléletidejű, instabil ncRNS-eket kerestek dohányban (*Nicotiana tabacum*), ekkor akadtak rá először a GUT15 ncRNS-re (NtGUT15). Az NtGUT15 RNS két splice variánsa 1.7 és 1.9 kilobázis (kb) hosszú (Taylor & Green, 1995). Később *A. thaliana*-ban is azonosították (AtGUT15), illetve megmutatták, hogy tartalmaz egy rövid szakaszt, mely erős konzerváltságot mutat a dohány homológgal, és mikro-peptid (75-78 aminosav) kódolásra is képes lehet (van Hoof és mtsai., 1997).

2016-ban Hsu és munkatársai (Hsu és mtsai., 2016) Ribo-Seq módszerrel azonosítottak egy 37 aminosav (as) hosszúságú peptidet *Arabidopsis* gyökerekben, mely a GUT15 génre/transzkriptumra illeszthető. Ez a peptid azonban nem azonos a GUT15-ben fellelhető leghosszabb, korábban megfigyelt/prediktált, 75 as hosszú peptiddel, funkciója egyelőre ismeretlen. Feltételezhető, hogy az AtGUT15 lncRNS riboszóma-szkennelés következtében degradálódik (ahogy azt más rendszerekben is felvettették, (Carlevaro-Fita és mtsai., 2016), mely folyamat hozzájárulhat a rövid félélet-idő szabályozásához.

A későbbi, dohányon végzett vizsgálatok kimutatták, hogy a NtGUT15 stabilitása szövet-specifikus, pollen sejtekben markánsan megemelkedik (Ylstra & McCormick, 1999).

A valós szerep megállapítása érdekében a *Tsuji labor* citokinin hormonra érzékeny géneket vizsgált uborkában (*Cucumis sativus*). Ezek között volt a CR20 (CsCR20); megállapították, hogy citokinin

hormon-kezelés hatására a CsCR20 transzkriptum szintje csökken, de halmozódik sebzéskor, sötétben és bizonyos szövet típusokban (gyökér, felnőtt és öregedő levelekben)(Teramoto és mtsai., 1995). Azt is kimutatták, hogy a CsCR20 és *Arabidopsis* homológja (*AtCR20*) tartalmaznak egy konzervált régiót (ami a GUT15-ben is jelen van), mely másodlagos szerkezet kialakítására alkalmas (Teramoto és mtsai., 1996).

Végül a *Jarmolowski labor* vizsgálta az *A. thaliana* GUT15 lncRNS gént és termékét; megállapították, hogy az RNS Polimeráz II (Pol II) által íródik át, 5'-CAP strukturát hordoz és 3'-poli-adenilált végű; továbbá, hogy a GUT15 naszcens RNS utolsó intronja prekurzorként szolgál két tRNS-szerű molekula éréséhez, de ehhez biológiai relevanciát nem tudtak társítani (Plewka és mtsai., 2018).

Összegezve tehát, annak ellenére, hogy évtizedek óta, több kutatócsoport munkája nyomán számos adat kiderült a GUT15/CR20 gén családról (hormonális szabályozás, sötét és hő-indukció, szövet-specifikus stabilizáció/destabilizáció), meglepő módon a funkciójuk továbbra is ismeretlen, valódi szerepükről mindezidáig csak hipotézisek és feltételezések születtek.

3.2 lncRNS-ek vizsgálati lehetőségei

Általánosságban az lncRNS transzkriptumok abundanciája egy nagyságrenddel alacsonyabb az mRNS-ekhez viszonyítva, emiatt nehezebb vizsgálni ezeket a molekulákat és funkcióikat **genetikai és molekuláris biológiai** (*wet labor*) módszerekkel. Bár más okok miatt, de az lncRNS gének **bioinformatikai** (*in silico*) vizsgálata is sokkal nehezebb. Az lncRNS génekben nem fellelhető a *start* kodon (ATG), a *stop* kodon (TGA, TAG vagy TAA), nincs bennük *nyílt leolvasási keret* (open reading frame, ORF), a *GC arányuk* is tágabb határokon belül mozog (az mRNS-ek GC aránya közelít az 50%-hoz, hiszen elméletben az összes aminosavat kódolniuk kell, ami nem feltétel az lncRNS-ek esetében). Emellett ezek a gének nem konzerváltak, tehát a más fajokból származó információ nem, vagy csak korlátozott mértékben hasznosítható. Mindezek figyelembevételével azonban, mégiscsak van néhány megközelítés, mely fehérje-kódoló gének vizsgálatánál használatos, de az lncRNS gének vizsgálatára is optimalizálható. Ezek mellett vannak olyan speciális módszerek, melyek lncRNS gének/transzkriptumok *in silico* vizsgálatára speciálisan fejlesztettek ki.

3.2.1. Szekvenciaillesztés

A “Blaszt” vagy pontosabban a “nukleotid blaszt” módszer a DNS vagy RNS poli-nukleotid láncban a monomer szerkezeti egységeinek (nukleotidoknak) egymásutánosságát, és az egymásutánosság hasonlóságának mértékét (a konzerváltságát) vizsgálja két vagy több polimér esetén. Ez a módszer alkalmas a részleges vagy nagymértékű hasonlóságot mutató szekvenciák összehasonlítására.

3.2.2. Filogenetikai elemzés

A filogenetikai elemzés lehetővé teszi a vizsgált fajokban fellelhető homológ gének (transzkriptek, fehérjék) közötti evolúciós kapcsolatok felderítését. Elméleti alapja az, hogy egy csoport (klád) tagjai azonos evolúciós történéseken mentek keresztül, ezáltal hasonlóbbak egymáshoz, mint a többi, a kládba nem tartozó génhez/transzkripthez/fehérjéhez), ugyanakkor olyan egyedi jellemzőkkel is bírhatnak, amik a távoli ősökben nem voltak jelen. A kladisztika három alapfeltevésen nyugszik: az összes élőlény egy közös őstől származik (ez egyben az evolúció alapelve is), elágazó mintázatot hoz létre és a tulajdonságokban idővel változás következik be a különböző vonalakban.

A filogenetikai kapcsolatokat filogenetikai fákon ábrázolják. A filogenetikai fákon az egyes ágak jelölik az egyes fajokat/géneket, ezek pedig kládokat alkotnak. Egy kládba egy faj/géncsoport legközelebbi közös őse, és azok leszármazottai tartoznak.

A fajok közötti rokonsági kapcsolatok vizsgálatához szekvenciaillesztést végezhetünk a fajok vizsgálni kívánt szekvenciái között és ezek hasonlósága alapján tudunk következtetni a fajok közötti hasonlóságra. A technológia fejlődésének köszönhetően mára bőven rendelkezésre állnak a gyors és felhasználóbarát szoftverek a szekvenciaillesztés elvégzésére, mint például a Clustal, vagy a Malign (Hickson és mtsai., 2000).

A többszörös szekvenciaillesztések során a szoftverek meghatároznak egy közös „ős-szekvenciát”, amiből a vizsgált szekvenciák sokszoros evolúciós történések során kialakulhattak. Ennek eredményeképp a vizsgált szekvenciák minden egyes bázisa valamilyen módon az ősszekvenciából származtatható. Mivel a különböző változások a szekvenciákban különböző mértékű evolúciós változásokat eredményezhetnek, ezek vizsgálatával meghatározható a szekvenciák közötti rokonság fok. Például a fehérje kódoló gének esetén a kodonok harmadik

tagjának (“lőtyögő nukleotid”) megváltozása sok esetben nem okoz aminosav változást, míg a tripletten belül lévő első két nukleotid változása többségében igen. Ebből következik, hogy a megváltozott szekvenciák pozíciójából adódóan egy változás kisebb és nagyobb evolúciós eseményt is jelölhet (Phillips és mtsai., 2000).

Bár a fent említett, és hasonló szekvencia-illesztő szoftverek használata nagy kényelemmel jár, könnyen torz eredményeket kaphatunk, ami miatt sok esetben szükséges az utólagos kézi korrigálás (Horner & Pesole, 2004). Az RNS-ek vizsgálata külön nehézségekbe ütközik (Fallmann és mtsai., 2017): míg a fehérje szekvenciák már 30% hasonlóság felett aránylag jól illeszthetők, az RNS-eknél a pontos illesztéshez legalább 60%-ra van szükség. Ennek oka a kisebb variálhatóság (4-féle nukleotid vs 22-féle aminosav) és az, hogy különböző elsődleges szerkezettel (szekvenciával) rendelkező RNS-ek is hasonló másodlagos szerkezetet alakíthatnak ki (Bussotti és mtsai., 2013).

3.2.3. K-mer analízis

A fehérjékkel szemben, ahol szoros kapcsolat áll fenn a szekvencia és a funkció között, ezáltal a homológ szekvenciával rendelkező fehérjék funkciója is hasonló lehet, az lncRNS-eknél a szekvencia maga nem szolgál ilyen erős információval. Mindemellett, a homológ lncRNS-ek funkciójának ismerete (más organizmusban) sem elég önmagában a vizsgált lncRNS funkciójának megértéséhez.

Ezek következtében az lncRNS-ek vizsgálata eltérő technikákat igényel. Mivel az lncRNS-ek funkcióját feltehetőleg a hozzájuk kapcsolódó RNS-kötő fehérje partnerek határozzák meg, ezek pedig rövid, 3-8 bp hosszú szakaszokhoz (*cis*z elemekhez, ún. K-merekhez) kötődnek, az azonos/hasonló funkcióval rendelkező RNS-eknek tartalmazniuk kell az adott fehérje- (partner-) kötőhelyeket. Több esetben azonban megfigyelték, hogy az azonos funkcióval rendelkező lncRNS-ek egészüket tekintve nem rendelkeznek szekvencia homológiával, ami alapján arra lehet következtetni, hogy bár tartalmazzák a kötőhelyeket, nem szükséges, hogy a kötőhelyet kialakító nukleotidok azonos sorrendben szerepeljenek, ennek következtében a szekvencia homológia sem lesz kimutatható.

A K-mer analízis során ennek megfelelően az lncRNS-ekről egy K-mer ujjlenyomat készül, a különböző K-mer hosszúságok alapján (pl. 3-6 nt hosszú K-mer) és ezek kerülnek összehasonlításra. Az ujjlenyomatok kalkulálása során kiszámításra kerül a megadott hosszúságú

szakaszra az összes lehetséges K-mer szekvencia variáció gyakorisága és ezek összehasonlításával kaphatjuk meg a teljes szekvenciák közötti hasonlóság mértékét (Kirk és mtsai., 2018).

4. Anyag és módszer

4.1 Számítógépes környezet

Munkánk során Windows 10 operációs rendszerben és Linux környezetben (Linux WSL) dolgoztunk, illetve amennyiben szükség volt rá, VirtualBox-on keresztül futtatott Linux Ubuntu 20.04 operációs rendszerben. A Python programokat Anaconda navigator felületen keresztül Jupyter Notebookban írtuk és futtattuk és Python 3.8 verziót használtunk hozzá.

4.2. Szekvenciaillesztés

A szekvenciaillesztést elvégeztük lokálisan és online is. A lokális szekvenciaillesztést az Ensembl Plants ([Ensembl Plants](#)) adatbázisában található genomi, cDNS és ncRNS szekvenciák az *AtGUT15* (AT2G18440) konzervált régiójának első 37 bázis hosszú szekvenciája alapján végeztük (GACCTTTGCCATGTCAGGTGCGCTTGCATGGCAGGTC). A szekvenciaillesztést az NCBI Blast+ alkalmazás segítségével végeztük és egy Python programnyelven általunk írt programmal automatizáltuk. Az NCBI Blast+ beállításait az NCBI Blast weboldaláról ([Nucleotide BLAST: Search nucleotide databases using a nucleotide query \(nih.gov\)](#)) importáltuk, az eredetileg megadott beállításoktól, csak a „Program Selection” szekcióban a „Somewhat similar sequences (blastn)” opcióval tértünk el. Az NCBI Blast+ a szekvenciaillesztés eredményét egy .xml fájlban adta meg, amit az általunk írt program dolgozott fel. A program a blast lefuttatása után megnyitotta a .xml eredmény-fájlokat és az azokban található keresési eredményekkel dolgozott tovább, amennyiben azok *e* szignifikancia értéke a határérték (< 0.01) alatt volt. Minden találat esetén azonosította a szekvenciát tartalmazó fájlt, megkereste benne a szekvencia kezdőpontját és kivágta a kezdete előtti 50 bázistól a kezdete utáni 230 bázisig terjedő szakaszt. A szekvenciák tehát tartalmaztak egy 50 nt hosszú upstream szakaszt, a 120 nt hosszú HCR régiót (3. ábra), kb 50 nt GT-gazdag régiót, illetve 10 nt downstream szakaszt. Az így kivágott szekvenciákat a program fajnévvel, illetve a kezdő – és végponttal FASTA formátumban kiírta egy eredmény fájlba.

A cDNS és ncRNS szekvenciák esetén a további vizsgálatok megkönnyítése érdekében elválasztottuk az azonosított HCR előtti és utáni szakaszt, illetve magát a HCR-t és opcionálissá tettük, hogy ezek közül melyek kerüljenek az eredményfájlba.

Az online szekvenciaillesztést fajonként végeztük el a korábban megadott GUT15 szekvencia szakasszal. Ehhez a Biopython NCBIWWW és NCBIXML csomagokat használtuk és a korábbiakhoz hasonlóan Python programnyelven automatizáltuk. Az eredmények a GUT15 teljes HCR régiójához való hasonlóságát a Jalview szoftverrel határoztuk meg.

4.3. Filogenetikai elemzés

A szekvenciaillesztés eredményeit összegeztük és minden fajra létrehoztunk egy konszenzus szekvenciát az EMBOSS Cons online tool ([EMBOSS Cons < Multiple Sequence Alignment < EMBL-EBI](#)) segítségével. Ezekből a szekvenciákból végeztük el a szekvenciaillesztést.

Mivel az lncRNS-ek eltérő tulajdonságaiból adódóan nem vizsgálhatók pontosan a fehérjék elemzésére létrehozott programokkal, Erik S. Wright 2020-as cikke alapján (Wright, 2020) a szekvenciaillesztést a MAFFT online felületén ([MAFFT alignment and NJ / UPGMA phylogeny \(cbrc.jp\)](#)) végeztük el. Bár elsősorban fehérje szekvenciák vizsgálatára lett fejlesztve, nagy pontossággal végez szekvenciaillesztést az (lnc)RNS-ek esetén is.

A vizsgált fajokból a filogenetikai fát az NCBI Taxonomy azonosítójuk alapján az NCBI online eszközével készítettük el ([Common Taxonomy Tree \(nih.gov\)](#)).

Az eredményeket mindkét esetben a MAFFT által ajánlott Phylo.io-val ábrázoltuk.

A filogenetikai fákat az R ggplot2, ggtree, cowplot, seqinr, ade4 és ape csomagok használatával hasonlítottuk össze.

4.4. Másodlagos szerkezet meghatározása

Az lncRNS-ek másodlagos és harmadlagos szerkezetének meghatározását a cDNS/ncRNS mintákon, a legszűkebb értelemben vett, 120 bázis hosszú HCR régióon végeztük el. Ehhez több programot használtunk, az Mfold ([www.mfold.org](#)), McGenus ([McGenus: A stochastic algorithm for the prediction of RNAs secondary structures with pseudoknots \(ipht.fr\)](#)), Kinefold ([KineFold website \(curie.fr\)](#)), RNA Composer ([RNAComposer \(put.poznan.pl\)](#)) és az RNAfold ([RNAfold web server \(univie.ac.at\)](#)) programokat.

4.5. K-mer analízis

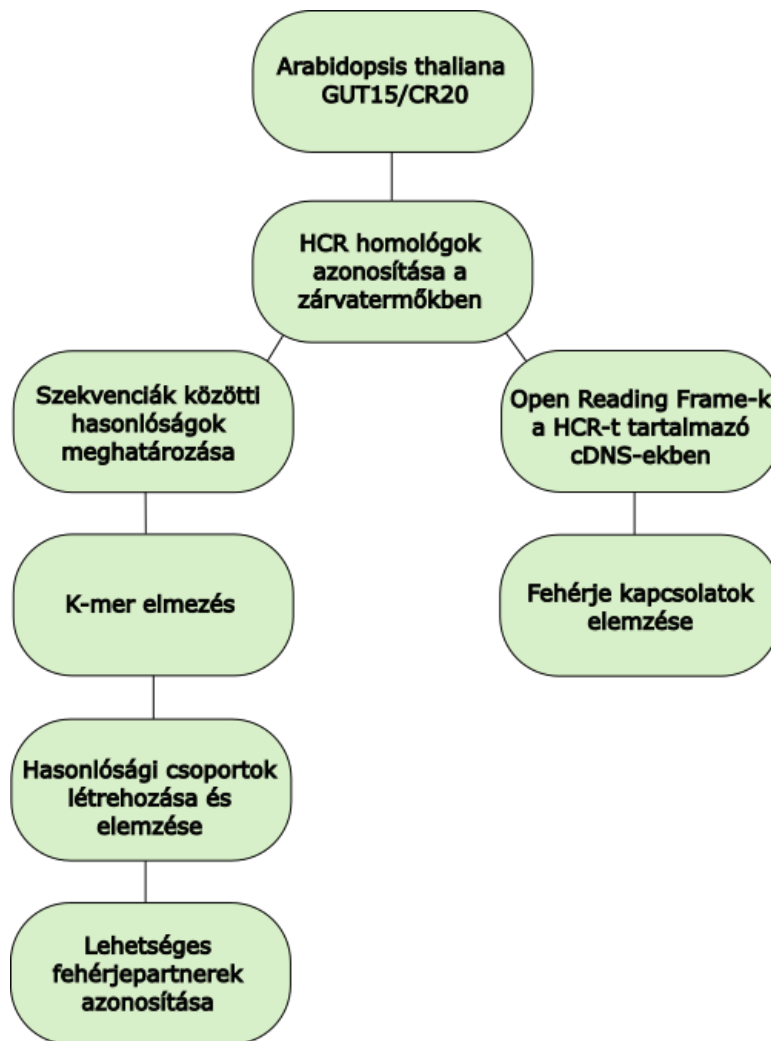
A K-mer analízist a cDNS és ncRNS szekvenciákon végeztük el. Ehhez a SEEKR oldalát ([SEEKR \(https://app.med.unc.edu/seekr/home\)](https://app.med.unc.edu/seekr/home)), illetve a Github repository ([CalabreseLab/seekr: A library for counting small kmer frequencies in nucleotide sequences. \(github.com\)](https://github.com/CalabreseLab/seekr)) leírása alapján a SEEKR Python könyvtárat használtuk, amit PyPI repository-ból töltöttünk le és Linux Ubuntu 20.04 operációs rendszeren parancssorban futtattunk.

4.6. Hierarchikus klaszterelemzés

A K-mer analízis során elvégzett Pearson-korreláció eredményeit .csv fájlként felhasználva végeztük el a hierarchikus klaszterezést, az R „amap” csomag használatával. Az eredményt a „ggplot” és „phytools” csomagok segítségével ábráztuk.

4.7. ORF-ek azonosítása

Az ORF-ek keresésére az Orfipy Python könyvtárat használtuk, amit a PyPI repository-ból töltöttünk le és Linux Ubuntu 20.04 operációs rendszerben parancssorok segítségével futtattunk, illetve Python nyelven automatizáltunk.



2. ábra: a GUT15/CR20 lncRNS-ek vizsgálatának lépései

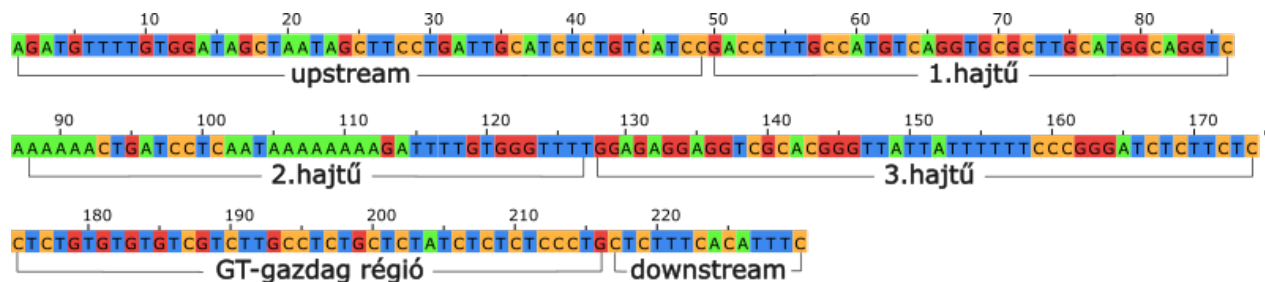
Vizsgálataink során az *Arabidopsis thaliana* GUT15 lncRNS konzervált HCR régiójának szekvenciájából indultunk ki. Szekvenciaillesztés segítségével meghatároztuk GUT15/CR20 homológokat az általunk használt adatbázisban elérhető zárvatermő szekvenciákban, majd vizsgáltuk hasonlóságukat, illetve illeszkedésüket a fajok filogenetikai fájára. Az azonosított DNS, cDNS és ncRNS szekvenciáknak meghatároztuk a K-mer ujjlenyomatát, ezek összehasonlításával pedig kialakítottunk hasonlósági csoportokat (klasztereket), ezeket pedig tovább elemeztük, kísérletet tettünk a feldúsult K-merek, illetve ennek segítségével a lehetséges fehérje partnerek azonosítására.

A cDNS szekvenciákban nyílt leolvasási kereteket (ORF) kerestünk és azonosítottuk az ezek szekvenciájára illeszkedő fehérjéket, majd ezeket tovább elemeztük.

5. Eredmények és megvitatásuk

5.1. Szekvenciaillesztés és filogenetikai elemzés

Munkánk során az *Arabidopsis* GUT15/CR20 család homológokból indultunk ki. A géncsaládba való tartozás egy magas konzerváltságot mutató szakasz (highly conserved region, HCR) alapján állapítható meg. A HCR régióban lévő első hajtű a leginkább konzervált (3. ábra). Elemzéseink első fázisában célunk volt (i) megállapítani, hány fajban és hány kópiában található meg a GUT15/CR20 konzervált HCR régiójának első hajtűje, (ii) milyen mértékű a konzerváltság az így azonosított szekvenciák – és azok teljes HCR hossznyi szekvenciája között, illetve (iii) ezek milyen módon illeszkednek a vizsgált növények taxonómiai fájára.



3. ábra: az *Arabidopsis thaliana* GUT15 (AT2G18440) általunk vizsgált szekvencia részlete

A GUT15/CR20 lncRNS-ek jellegzetes, másodlagos szerkezetet kialakítani képes konzervált HCR régióval rendelkeznek, amit három hajtű alkot és egy változó hosszúságú GT-gazdag régió követ. Az első hajtű ezek közül a leginkább konzervált, ezért ez alapján végeztük el a szekvenciaillesztést. Az eredmény szekvenciák az első hajtű kezdetétől upstream 50 nt-től indultak és tartalmazták a 3 hajtűt (120 nt), a GT-gazdag régiót (kb 50 nt) és kb 10 nt downstream szakaszt, így összesen 230 nt hosszúak voltak.

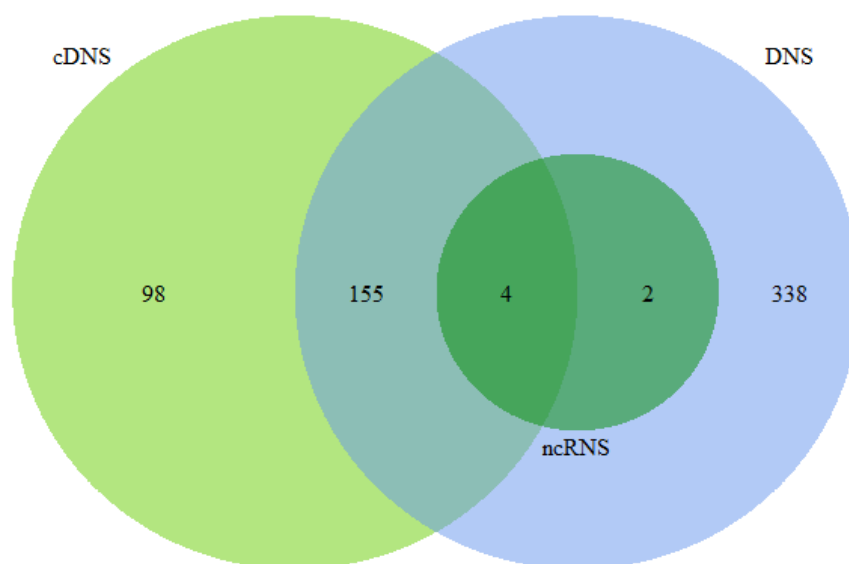
5.1.1 Szekvenciaillesztés

A szekvenciaillesztést az adatbázisokban elérhető összes faj szekvenciáin elvégeztük. Ez legfőképp zárvertermőket jelentett, ezeken kívül pedig alga és moha fajok adatait. Összesen 102 faj genomi (DNS adat), 105 faj transzkriptom (cDNS adat) és 57 faj lncRNS-ként annotált szekvenciáit vizsgáltuk.

Ennek eredményeképpen 92 faj genomi szekvenciájában azonosítottunk HCR régiót, 87 faj esetén legalább 2 kópiában volt fellelhető, így összesen 499 HCR-t tartalmazó DNS lókuszt

azonosítottunk. A cDNS szekvenciákban 54 fajban összesen 257 esetben, míg az ncRNS adatok alapján 4 fajban 6 HCR-t találtunk (4. ábra). A DNS adatok tanulmányozása során, 10 faj volt melyben nem találtuk meg a HCR szekvenciát, ezek a fajok majdnem kizárólag algák és mohák közé tartoztak (a fent említett 10 faj cDNS és ncRNS szekvenciáiban sem volt fellelhető a HCR szekvencia). Továbbá az adatbázisokban fellelhető nyitvatermő fajok szekvenciáiban sem található meg a HCR szekvencia (DNS és cDNA/RNS adatok alapján egyaránt, Szaker et al., *előkészületben*).

Az eredmények arra engednek következtetni, hogy a HCR régiót tartalmazó gének jelenléte a zárvatermő életmódhoz kapcsolódik (megj: A fehér tündérrózsa (*Nymphaea colorata*) genomi szekvenciáiban sem volt megtalálható a HCR, ami vagy a genomi szekvencia hiányossága miatt van, vagy esetleg egy utólagos génkiesés magyarázhatja).



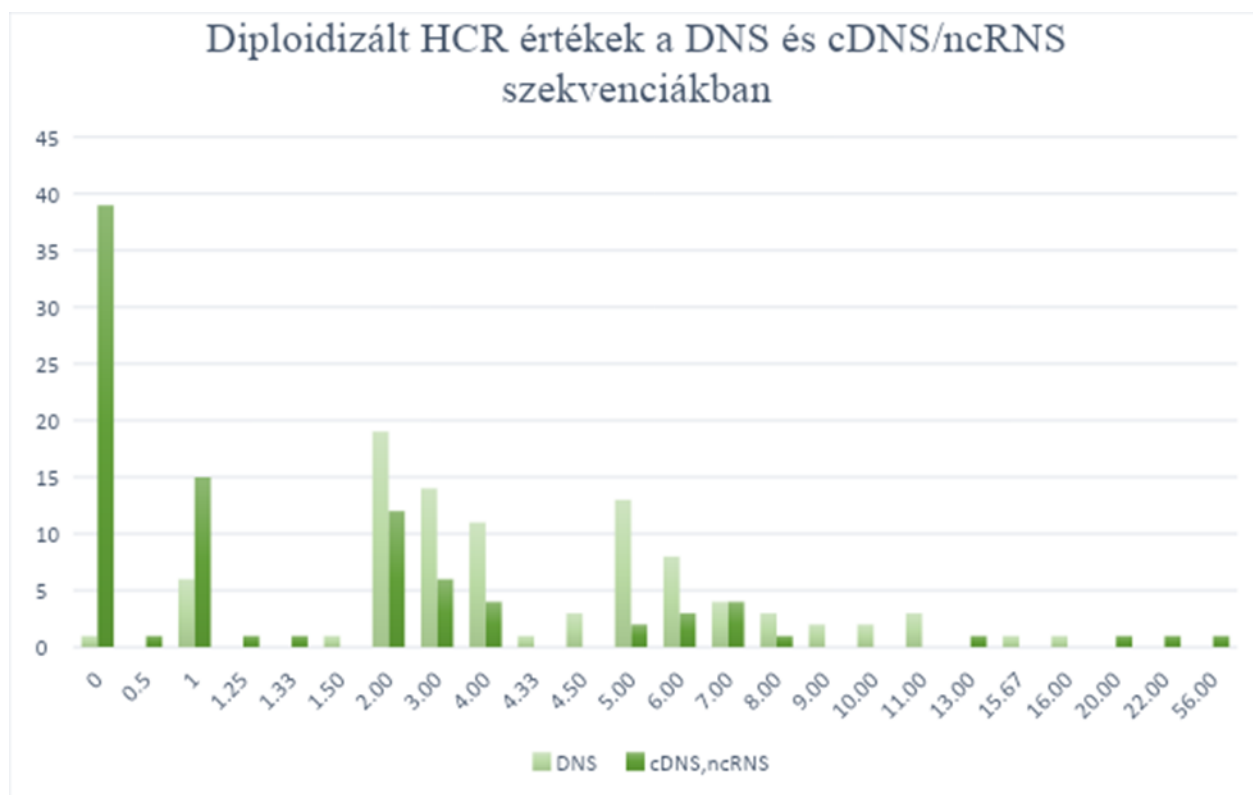
4. ábra: az azonosított HCR-ek mennyiségének összegzése

A szekvenciaillesztést 102 faj elérhető DNS, 105 faj cDNS és 57 faj ncRNS szekvenciáin végeztük el. Ennek eredményei alapján a DNS szekvenciákban 499, a cDNS-ekben 257 és az ncRNS szekvenciákban 6 HCR-t azonosítottunk. Ezeket fajonként összevetve összegeztük és azonosítottuk azokat az eredményeket, amik mind a genomi, mind a cDNS/ncRNS mintákban jelen voltak.

Mivel sok fajban több (legalább 2) kópiában van jelen a gén, felmerült, hogy ennek biológiai relevanciája van, például (i) a GUT15/CR20 lncRNS-ek dózisfüggően működhetnek, (ii) interakció van a gének vagy termékeik között, és (iii) egy multimer komplex funkcionalitásához mindkét lncRNS egyidejű jelenléte szükséges.

A növényekre jellemző poliploidizáció rokon fajok között jelentős eltéréseket okozhat az egyes gének kópiaszámában. Annak érdekében, hogy összehasonlíthatóvá tegyük az egyes fajokban talált HCR számokat, kiszámoltuk a HCR diploid értéket: összesítettük fajonként külön-külön a genomi és cDNS/ncRNS találatokat, meghatároztuk a fajra jellemző ploidia fokot és – amennyiben nem diploid volt – elosztottuk azzal az értékkel, ahányszoros volt a ploidia a diploidhoz képest. Ez pl. egy hexaploid faj esetén 3-mal való osztást jelentett. Az így kapott „*diploid HCR érték*” tehát meghatározta azt, hogy $2n$ kromoszómakészlet esetén hány HCR-t tartalmazna az adott faj. A „diploidizáció” összesen 13 fajt érintett, köztük 3 *Triticum* és 2 *Brassica* fajt is.

A kapott eredmények azt mutatták, hogy a DNS szekvenciákban a HCR tartalmú génből jellemzően legalább kettő van jelen (6 fajt azonosítottunk 1 HCR-kópiaszámmal és 19-et kettővel), de a 3, 4 és 5 HCR-kópia számmal rendelkező fajok száma is jelentős (X. ábra). Míg a fajok többségében ezt a 2-5 kópiaszámot azonosítottuk (összesen 61 fajban), a diploid HCR érték egyes növényeknél kiemelkedő értékeket mutatott (pl. a *Dioscorea rotundata* és a *Camelina sativa* is 16 HCR kópiát kódol) (5. ábra).



5. ábra: A növényekben előforduló HCR tartalmú gének száma diploid genomra vonatkoztatva. Az X tengely a diploidizált HCR kópiaszámot, az Y tengely a fajok számát mutatja; DNS adatok (halványzöld), cDNS/RNS adatok (sötétzöld).

A cDNS/ncRNS szekvenciák vizsgálata során a diploid HCR értékek változatosabb képet mutattak, mint a DNS minták esetén. Míg a 0 és 1 kópiával rendelkező fajokból találtunk a legtöbbet, 20, 22 és 56 HCR-tartalmú szekvenciát is azonosítottunk a *Dioscorea rotundata* (56), *Brachypodium distachyon* (22) és *Aegilops tauschii* (20) fajokban. Emellett nem volt elhanyagolható a 2 és 3 kópiával rendelkező fajok száma sem.

A diploid HCR értékek mind a DNS, mind a cDNS/ncRNS adatsorok esetén rendkívül változatosak, ugyanakkor közöttük lényeges eltérés mutatkozik. Míg a genomi minták alapján egyértelműen megállapítható, hogy a fajok többségében legalább 2, vagy több HCR kópia van jelen, a cDNS-ek és ncRNS-ek adatokban nincs, vagy csak egy HCR homológ található meg fajonként. Az eltérés adódhat abból, hogy a DNS-ben jelenlévő HCR szekvenciák nem minden esetben íródnak át RNS-é, vagy az is elképzelhető, hogy a transzkriptom adatbázisok hiányosak az eltérést mutató fajokban.

Ezzel szemben 3 fajban a genomi DNS szekvenciáknál jóval magasabb számú HCR kópia van jelen a cDNS/ncRNS adatokban; elképzelhető, hogy a különböző RNS splice variáns annotációk

okozzák ezt a többletet (megj: *Arabidopsis*-ban is több splice variánsa van mind a GUT15, mind a CR20 gén transzkriptumainak, alátámasztva ezt a felvetést).

Ezek az eredmények arra engedtek következtetni, hogy a GUT15/CR20 gének vagy a róluk átíródó lncRNS termékek a zárvatermőkben általában több (leggyakrabban 2) kópiában szükségesek és mivel minden vizsgált zárvatermő csoportban jelen vannak (kivétel a tündérrózsa), feltehetően biológiai funkcióval is rendelkeznek.

Adataink felvetik a lehetőségét annak, hogy a GUT15/CR20 géncsaládon belül funkcionális alcsoportok alakultak ki, és talán emiatt van szükség két vagy több homológra; amennyiben ez igaz, az alcsoport speciális tulajdonságait a HCR régió kívüli szekvenciák határozhatják meg.

5.1.2 Filogenetikai fa készítése

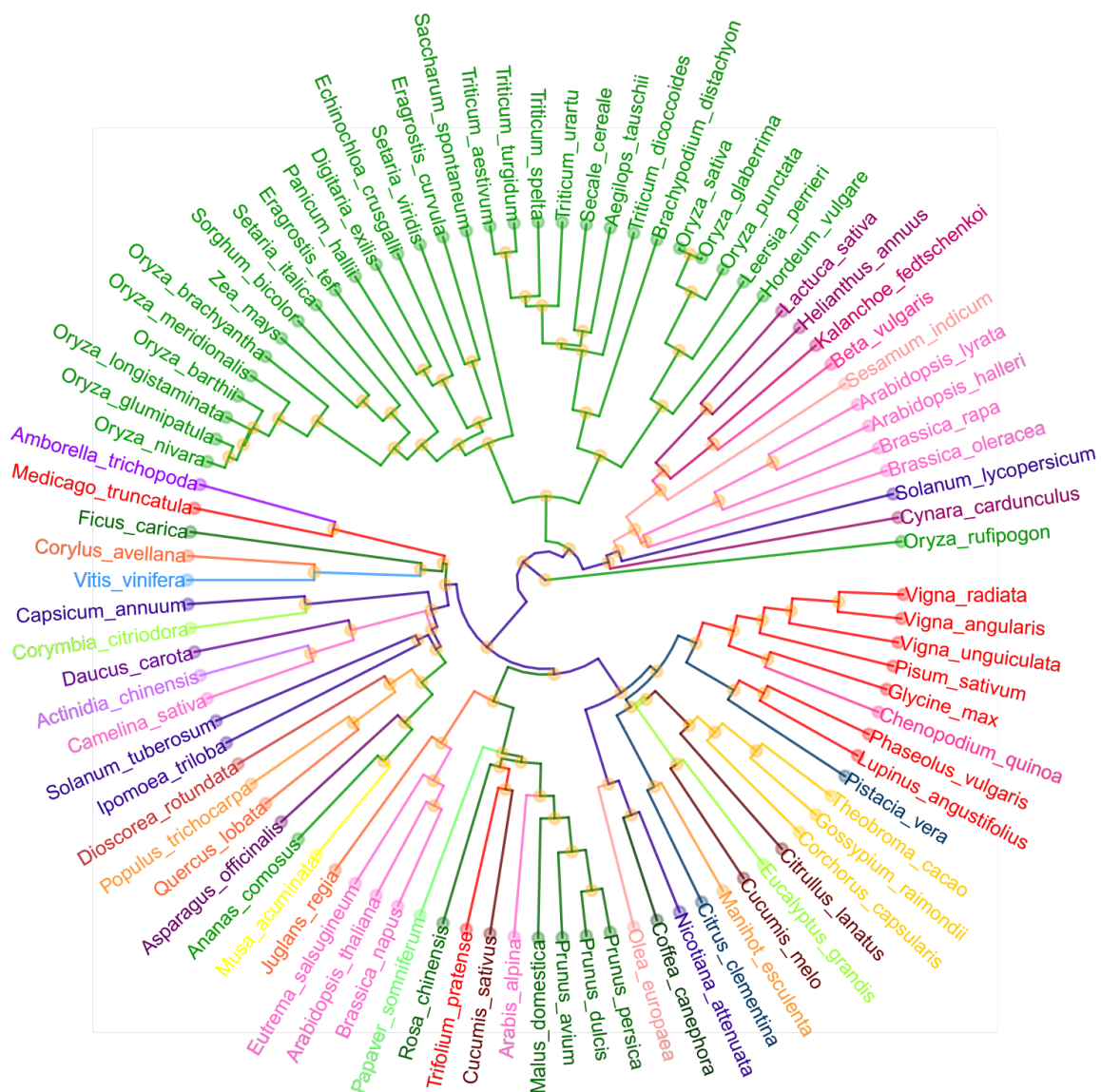
Hogy rátekintést nyerjünk a HCR régió változására/evolúciójára a törzsfajlás során létrehoztunk egy hasonlóságon alapuló HCR filogenetikai fát és ezt összevetettük a növényi törzsfajlás fával. Ehhez először a HCR-t tartalmazó szekvenciáknak létrehoztuk fajonként a konszenzus szekvenciáját (mind a DNS, mind a cDNS/ncRNS adatok felhasználásával). Ezeken a konszenzus szekvenciákon szekvenciaillesztést végeztünk és filogenetikai fán ábráztuk. A filogenetikai fát egyenként, illetve egymáshoz hasonlítva elemeztük.

A genomi szekvenciák illesztését ábrázoló filogenetikai fa két nagy csoportra osztható, ezekbe a csoportokba csak az *Oryza rufipogon* szekvenciája nem illeszkedik bele (6. ábra). A két csoport további 3-3 alcsoportra tagolódik. A fát a fajok alapján készített törzsfával összehasonlítva azt tapasztaltuk, hogy bár sok a különbség, a HCR-ek az egyes csoportokban jellemzően hasonlóan változnak az evolúció során és az egy taxonómiai csoportba tartozó fajok a legtöbb esetben megmaradnak egy alcsoporton belül. Kivételt képeznek a libatopfélék (*Chenopodiaceae*) rendjébe tartozó *Chenopodium quinoa* és *Beta vulgaris*, az ajakosvirágúak (*Lamiales*) rendjébe tartozó *Olea europaea* és *Sesamum indicum*, illetve a burgonyavirágúak (*Solanales*) rendjébe tartozó *Solanum lycopersicum*, *Solanum tuberosum*, *Nicotiana attenuata*, *Capsicum annuum* és *Ipomoea triloba* fajok, amik külön csoportokba kerülnek a filogenetikai fán. A legjelentősebb konzerváltság a perjevirágúak (*Poales*) rendjében fedezhető fel, de a rózsavirágúakban (*Rosales*) és a káposztafélékben (*Brassicaceae*) is jelentős.

Míg a fajok általában közeli rokonság mellett helyezkednek el, a rendek pozíciója nem egyezik a taxonómiai fán látható elrendezéssel. Például a taxonómiai fán a *Poales* és *Rosales* rendek egymás

mellett helyezkednek el, ezzel szemben a szekvenciaillesztés eredménye alapján jóval nagyobb távolság van közöttük. Ezzel szemben a *Brassicaceae* fajok HCR konzerváltsága jóval hasonlóbbnak bizonyul a *Poales* rend tagjaihoz, mint a taxonómiai fán a két növénycsoport.

Bár voltak eltérések a taxonómiai és a szekvenciák alapján elkészített filogenetikai fák között, összességében azt tapasztaltuk, hogy a HCR szekvenciák csoportosítása nagy vonalakban megfelel a taxonómiai csoportoknak. Ezt mutatja az is, hogy a jelentős eltérések jellemzően a kis, 2-5 fajszámmal rendelkező rendekben tapasztalhatók, míg a nagyobb létszámmal jelenlévő rendeknél a fajok pár kivételtől eltekintve egy és két csoportban maradtak (6. ábra).



Actinidiaceae	Chenopodiaceae	Lamiales	Poales
Amborellaceae	Crassulaceae	Malpighiales	Rosales
Apiaceae	Cucurbitaceae	Malvaceae	Rubiaceae
Asparagaceae	Dioscoreaceae	Musaceae	Sapindales
Asteraceae	Fabaceae	Myrtaceae	Solanales
Brassicaceae	Fagales	Papaveraceae	Vitaceae

6. ábra: a genomi HCR szekvenciák alapján készített filogenetikai fa

Az azonosított HCR-eknek fajonként meghatároztuk a konszenzus szekvenciáját, majd ez alapján többszörös szekvenciaillesztést végezve létrehoztuk a szekvenciák hasonlóságán alapuló filogenetikai fát. A filogenetikai fán a különböző színekkel jelöltük az eltérő nemzetségekbe tartozó fajokat. A filogenetikai fa alapján megállapítottuk, hogy az egyes eltérések ellenére, a HCR-ek konzerváltsága valószínűleg valamilyen mértékben követi a fajok evolúciós változásait.

Az elérhető cDNS/ncRNS adatokból létrehoztunk egy RNS-alapú HCR konszenzus szekvenciát és ebből készítettünk egy taxonómiai fát, amit összevetettünk az érintett fajok (tehát a cDNS szekvenciákban HCR-t tartalmazó fajok) törzsfajlódási fájával. A konszenzus szekvenciákon elvégzett szekvenciaillesztés eredményeit ábrázolva és összehasonlítva a taxonómiai, illetve a genomi szekvenciákból készített fával, a korábbiakhoz (a genomi/DNS alapú adathoz) hasonló eredményeket tapasztaltuk: az egyes taxonómiai csoportokba tartozó fajok (ahogy a DNS szekvenciák esetén is) egy/közel csoportban maradtak, míg a csoportok helyzete változott, mind a genomi szekvenciákból létrehozott filogenetikai fához képest.

Mivel a szekvenciaillesztés eredményeiből és a taxonómiai rokonságokból más szoftver segítségével készítettük el a filogenetikai fát, az egyes csoportokon belüli hasonlóságok valószínűleg nagyobb információ értékkel bírnak, mint maguknak a csoportoknak az elhelyezkedése.

A filogenetikai elemzések következő lépéseként a fajokra jellemző diploid HCR értékeket összevetettük a rendszertani elhelyezkedésükkel annak érdekében, hogy meghatározzuk, van-e kapcsolat a diploid HCR értékek és a rokonsági kapcsolatok között. Megállapítottuk, hogy a legtöbb esetben közel rokon fajoknál is lehet akár többszörös eltérés. Erre példák a *Poaceae* (perjefélék), vagy a *Brassicaceae* (káposztafélék) családok tagjai. Az egyes nemzetségek tagjaiban nagyobb hasonlóság volt felfedezhető, de ezek részletes kiértékeléséhez nem állt megfelelő mennyiségű adat a rendelkezésünkre.

Ezek az adatok arra engednek következtetni, hogy a HCR tartalmú gén már a zárvatermők korai őséiben megjelent, és szekvenciája erősen konzerválódott az evolúció során. Ez egy ősi funkcióra

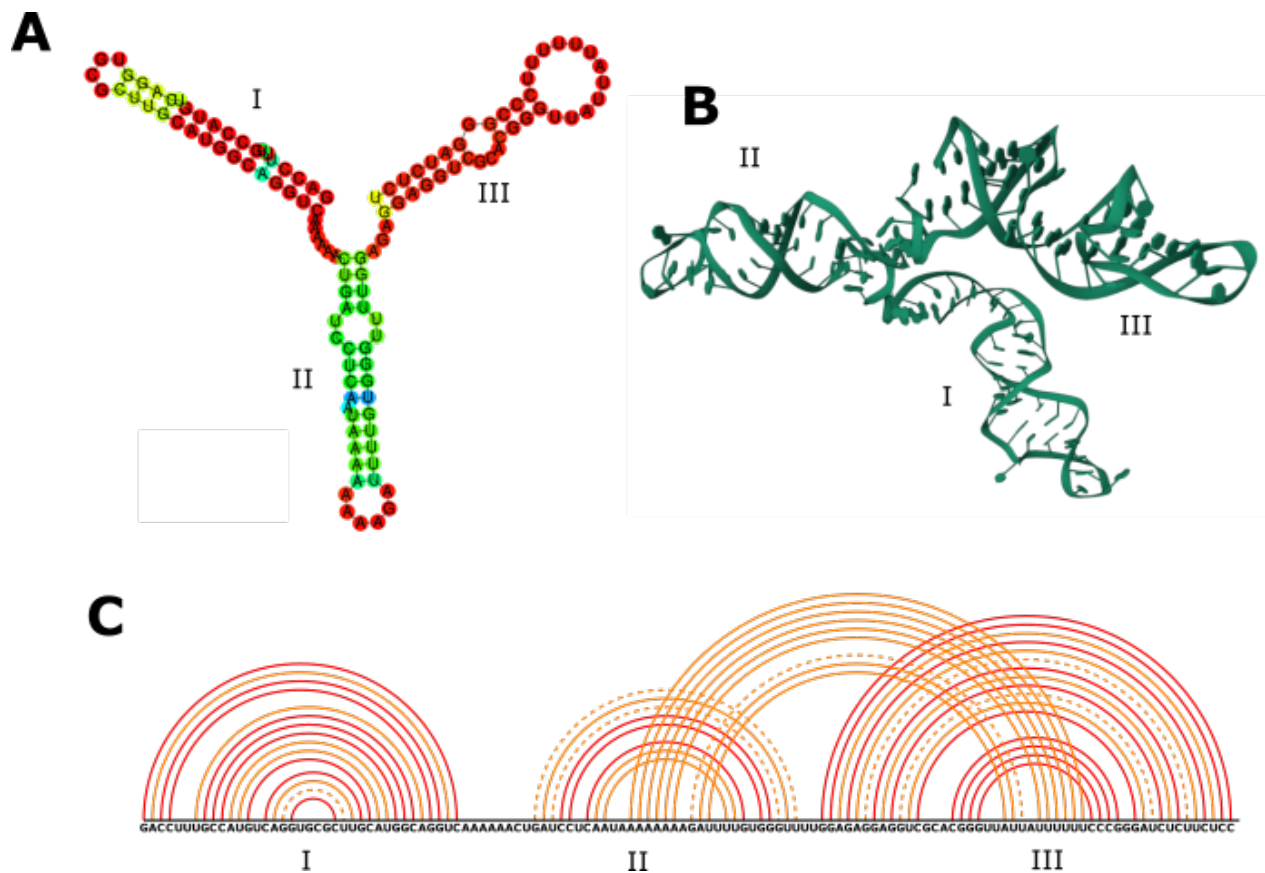
(biológiai relevanciára) utaló jel. Ezzel szemben néhány taxonómiai csoportban számuk változott, a leggyakrabban 2 kópia helyett megsokszorozódott pl. a *Poaceae* és *Brassicaceae* családokban, mely speciális funkciókra, finomszabályozásra ad lehetőséget. Más esetekben pedig, nem zárható ki, hogy a HCR tartalmú gének elvesztek az evolúció során, pl. a *Nymphaea colorata*-ban, vagy egyéb 1 kópiát tartalmazó fajokban (*megjegyzés*: az 1 kópiás vagy HCR nélküli fajok esetében, legvalószínűbb a technikai magyarázat, azaz, hogy az adatbázis hiányosság miatt nem találtuk meg; ugyanis, funkcióvesztés esetében az alacsony konzerváltságú HCR jelenlét lenne jellemző, de ilyen egy esetben sem találtunk).

5.1.3. A HCR régió szerkezeti elemzése

A GUT15/CR20 lncRNS géncsalád jellemző és feltételezhetően a funkcionalitást hordozó régiója a HCR domén (a konzervált ~120 nt). Hogy betekintést nyerjünk a struktúra és a funkció esetleges kapcsolatára, első lépésben prediktáltuk a másodlagos szerkezetét (120 nt hosszú HCR domén, az RNS upstream és downstream szakaszai nélkül) (3. ábra). Munkánkat az *Arabidopsis* GUT15 (AT2G18440) HCR szakaszának vizsgálatával kezdtük, majd ezt a szerkezetet összevetettük a más fajokban azonosított HCR-ek másodlagos szerkezetével, illetve a konszenzus HCR struktúrával. Az *AtGUT15* HCR egy hármas hajtú alakzatot hoz létre (3-féle prediktáló program nagyon hasonló szerkezeteket prediktált; ViennaRNAfold, McGenus, és RNA Composer, 7. ábra). A struktúra, és konzerváltságának vizsgálata során megállapítottuk, hogy az első hajtú a legkonzerváltabb, míg a második és harmadik kevésbé. A második hajtú duplaszálú régióját egy reverz-komplementer (többszöves) poli-A és egy poli-U nukleotid sorozat alakítja ki. A második és harmadik hajtú csúcsi részén nem-konzervált hosszúságú hurkok helyezkednek el; a csúcsi hurkok egy poli-A és egy poli-U (a McGenus predikcióban U helyett T szerepel) egymással képesek kölcsönhatni, ezáltal egy bonyolult háromdimenziós szerkezetet létrehozni, aminek esetleg funkcionális konzekvenciái lehetnek (*megj*: a hurkok közötti kölcsönhatás predikciójára csak a McGenus program alkalmas, 7C ábra).

Az eredmények alapján több feltételezésünk van a szerkezet-funkció kapcsolatra: (i) a három hajtú külön-külön partnerrel és/vagy funkcióval rendelkezik, (ii) az első hajtú és a második-harmadik hajtú együttese külön feladatot lát el, (iii) mindhárom hajtú együttműködésben lát el egy egységes funkciót, (iv) a középső poli-A/U hajtú egy hőérzékelő struktúra (alacsony olvadásponttal rendelkezik) mely “mozgatja” a kétoldalán lévő első illetve harmadik hajtú struktúrát és az

esetlegesen ehhez kötődő partnereket (ezt alátámasztja az a megfigyelés is hogy a GUT15/CR20 hőszabályozott expresszióval bír (Szaker et al., *előkészületben*)).



7. ábra: A GUT15 HCR régiójának in silico prediktált szerkezete

A GUT15 konzervált HCR régiója jellegzetes, 3 hajtúval rendelkező másodlagos szerkezetet alakít ki. Ezt többféle szoftverrel is meghatároztuk; (A) ViennaRNAfold, (B) RNA composer, (C) McGenus.

5.2. K-mer analízis, hierarchikus klaszterelemzés

A K-mer analízist két céllal végeztük el; egyrészt, hogy megvizsgáljuk (i) az egyes fajok HCR régiót tartalmazó génjeinek HCR és környékbeli szakaszai közötti (HCR +50nt up- és 60 nt downstream) hasonlóságát a DNS és RNS adatok alapján egyaránt, (ii) a homológok HCR régióin kívül eső szakaszok hasonlóságát RNS adatok alapján, és (iii) amennyiben vannak hasonlóságok akkor azonosítsuk hasonlósági csoportokat. Másrészt pedig, hogy egy fajon belül összehasonlítsuk

a GUT15/CR20 lncRNS-ek K-mer mintázatát más lncRNS-ek mintázatával, és ez alapján funkcionális információt nyerjünk (amennyiben a más lncRNS-ek jellemzettek). A GUT15/CR20 lncRNS-ek több funkcióval rendelkezhetnek; mivel a funkciót a fehérjepartnerek, a fehérjepartnerek kötődését az lncRNS-ben lévő kötőhelyek, azaz K-merek határozhatják meg, a K-mer analízis segítségével elméletben elkülöníthetjük a különböző lncRNS funkciókat.

A K-mer analízist 3-6 nt hosszúságú értékkel végeztük el. Az eredményeket két formában kaptuk meg, egyrészt a Pearson-korreláció eredményét, ami az egyes minták egymáshoz hasonlított K-mer ujjlenyomatát tartalmazta, másrészt pedig mintánként a hossz alapján normalizált K-mer értékeket. A hierarchikus klaszter analízishez a Pearson – korrelációk eredményeit használtuk és dendogram, illetve a mintákat klaszterekbe rendezett táblázat formájában elemeztük.

5.2.1. A HCR régió K-mer tartalmának vizsgálata

Bár a K-mer analízis elsődlegesen az lncRNS-ek vizsgálatára alkalmas, a K-mer ujjlenyomat fontos információkkal szolgálhat az egyes szekvenciák hasonlóságainak (3-6 nt hosszúságú, azonos összetételű bázistartalmú szakaszok), vagy éppen különbözőségeinek meghatározásához a DNS minták esetén is, főleg egy erősen konzervált szakasz esetén.

A DNS HCR régió vizsgálatának eredményei a genomi szekvenciák esetén az összes K érték esetén azt tükrözték, hogy bár a fajok között jellemzően nagyobb a hasonlóság, azokon a fajokon belül, ahol több szekvenciát is vizsgáltunk, a legtöbb esetben több hasonlósági csoport is elkülöníthető volt, mind a fajon belül, mind a többi fajhoz képest is. Ez a mintázat legjobban a *Triticum* fajoknál (*Triticum aestivum*, *dicoccoides*, *spelta*, *turgidum*, *urartu*) volt megfigyelhető. Itt a különböző kromoszómákon található szakaszok a többi faj azonos kromoszómáján lévőkhöz jobban hasonlítottak, mint az azonos fajon belül másik kromoszómán lévőkhöz (8. ábra).

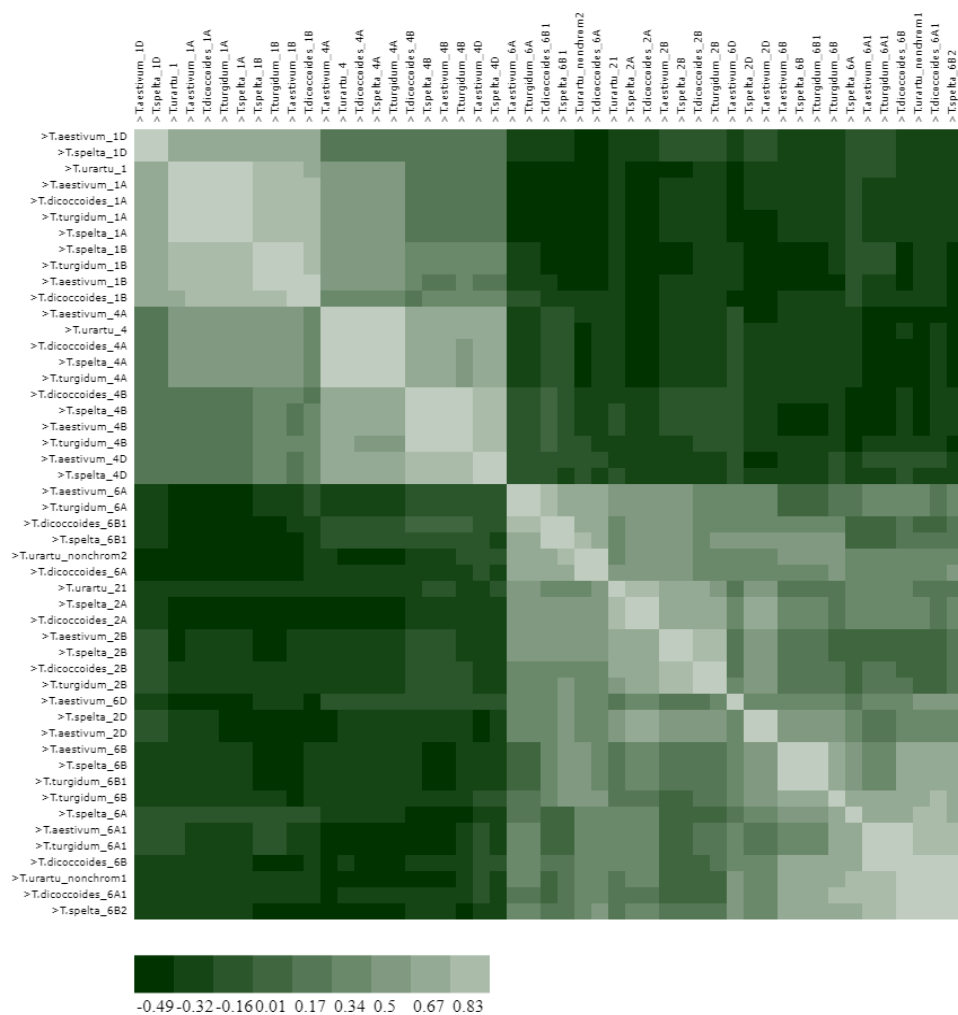
A cDNS/ncRNS HCR szekvenciák esetén a K-mer analízist duplikátumok nélkül, összesen 142 szekvencián végeztük el, a korábbiakhoz hasonlóan 3-6 K-mer értékekkel. Az eredmények megfeleltek a DNS szekvenciáknál tapasztaltaknak, azonban – mivel kevesebb faj kevesebb szekvenciáján végeztük el – kisebb mértékben mutatták a korábban tapasztalt mintázatokat. Mivel a cDNS/ncRNS szekvenciák esetén jóval kevesebb eredmény állt rendelkezésünkre, ezekből mindössze annyi következtetést tudtunk levonni, hogy a DNS kmer vizsgálatban tapasztalt mintázat itt is jelen volt, tehát a fajokon belüli feltételezhető csoportok a transzkripció, illetve a poszt-transzkripciós módosítások (pl. RNS splicing) során is megmaradtak.

5.2.2. *Triticum* fajok K-mer analízise és hierarchikus klaszterelemzése

A DNS mintákban, azokon belül is legjelentősebb mértékben a *Triticum* fajoknál látott mintázat majdnem az összes olyan faj, vagy rokon fajok esetén megfigyelhető volt, ahol nagyobb számú HCR szekvenciát azonosítottunk. Ez, bár kevésbé volt szembetűnő, szintén jelen volt a cDNS/ncRNS szekvenciák K-mer analízisének eredményeiben.

Ahhoz, hogy ezt a mintázatot részletesebben megvizsgáljuk, kiválasztottuk a fent is említett *Triticum* fajokat és ezek DNS és cDNS/ncRNS szekvenciáit együtt elemeztük tovább. Az első eredmények alapján azonban egyértelművé vált, hogy a DNS és cDNS/ncRNS szekvenciák között jelentős átfedés van, tehát a cDNS/ncRNS adatokban is azonosítottuk ugyan azt a HCR-t, mint a DNS szekvenciákban. Mivel a hasonlósági csoportok kialakítása szempontjából ezek a szekvenciák irrelevánsak voltak, eltávolítottuk a duplikátumokat és összesen 49 *Triticum* szekvenciával dolgoztunk tovább.

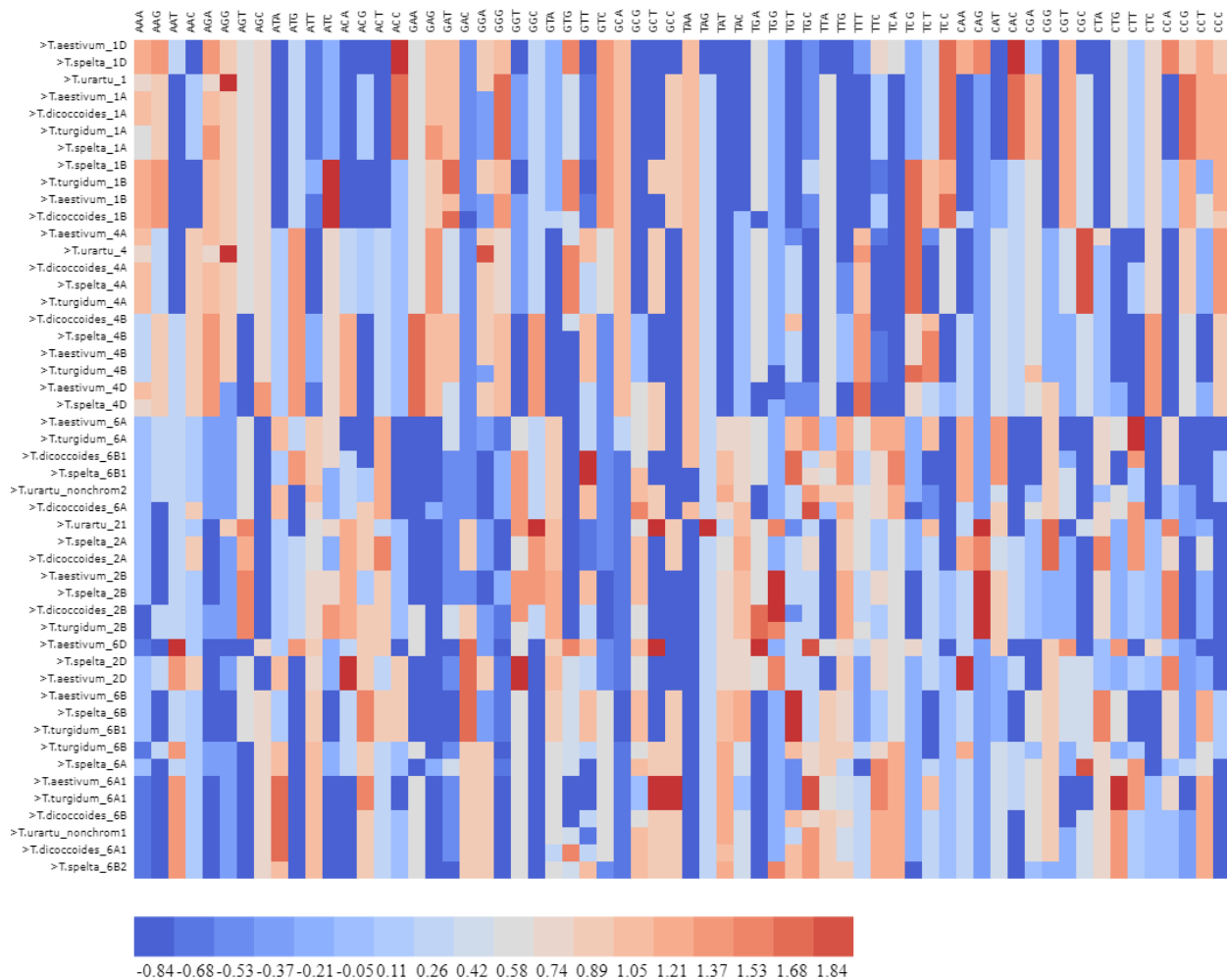
Az eredmények a K-mer méret függvényében különbözőképpen alakultak. Általánosságban az volt megfigyelhető, hogy a csoportok élesen elváltak egymástól, és az egyes kromoszómákon (1, 2, 4, 6 kromoszómákon) lévő HCR-ek jobban hasonlítottak egymáshoz a nemzetségen belül, mint a fajon belüli más kromoszómán lévő HCR-hez.



8. ábra: a *Triticum* fajok HCR régiójának Pearson-korrelációs mátrixa (K-mer = 3)

A *Triticum* fajok HCR-t tartalmazó DNS és cDNS szekvenciáit összesítettük és elvégeztük rajta a K-mer analízist, hogy különböző hasonlósági csoportokat különítsünk el. A Pearson-korrelációs mátrix alapján megfigyelhető, hogy az egyes kromoszómákon (1, 2, 4, 6) található HCR-ek jobban hasonlítanak egymáshoz, mint a fajon belüli, másik kromoszómán található HCR-t tartalmazó szekvenciákhoz.

A K-mer értékek alapján is megfigyelhető, hogy a szekvenciák több, nem a fajok alapján kialakult csoportra oszlanak és ezek K-mer értékei jelentősen különböznek egymástól, azonban míg a Pearson – korreláció eredményei alapján nagy vonalakban meghatározhatóak voltak az egyes csoportok, a K-mer profilok esetén ezek nem váltak el olyan tisztán. Azonban megfigyelhetők egyértelmű eltérések, mint például az egyik csoportban az „ACG” és „GTA” K-merek mennyisége volt kimagasló, egy másik csoportban a „TAT” és „TAC” K-mereké (8.,9. ábra).

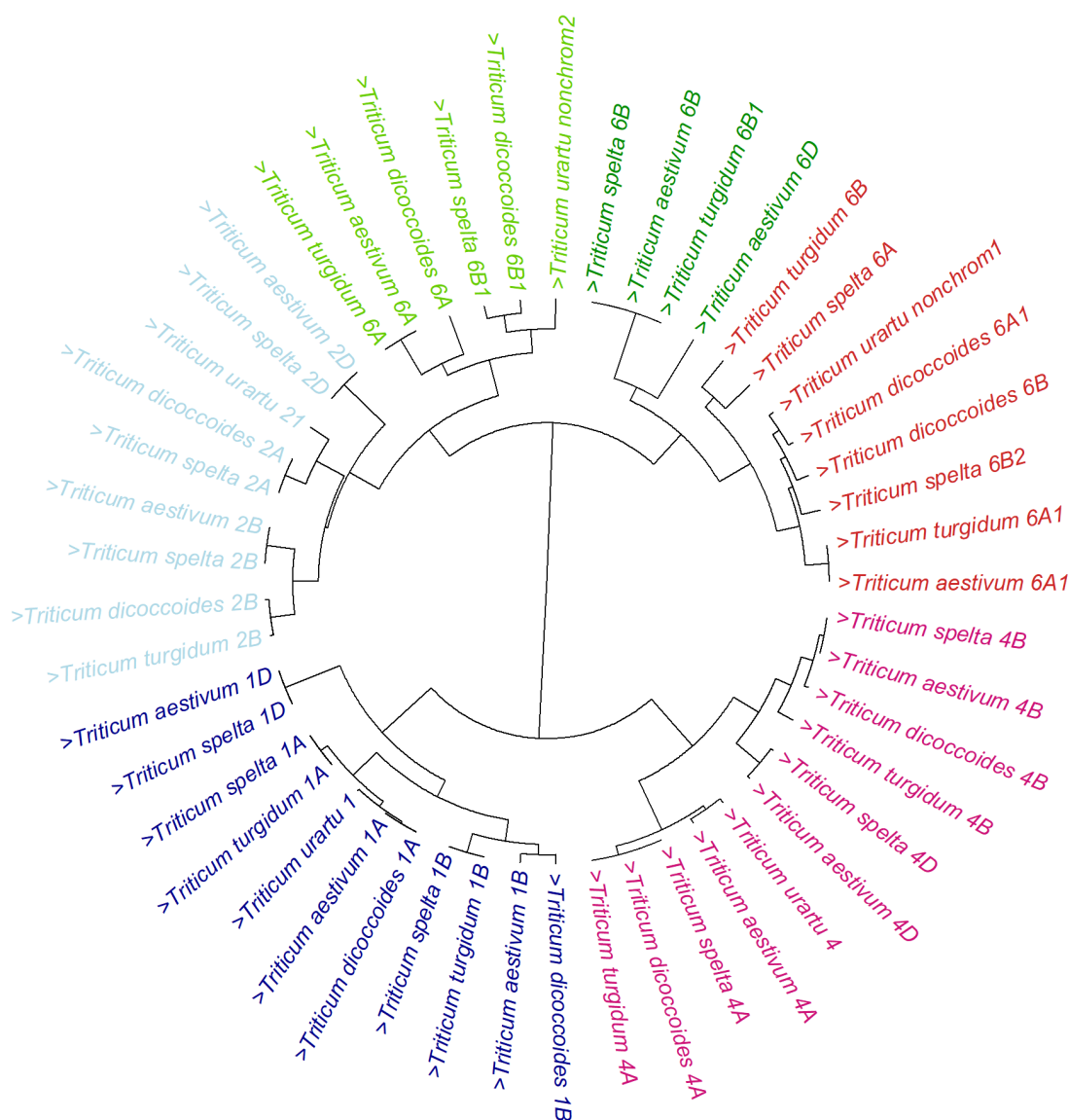


9. ábra: a *Triticum* fajok HCR régiójának K-mer mátrixa (K-mer = 3)

A *Triticum* fajok K-mer analízisének K-mer mátrixa alapján is megállapítható, hogy több csoport különül el, de nem olyan élesen, mint a Pearson - korrelációs mátrix esetén. Megfigyelhetők azonban jelentős eltérések mindkét csoportban, mint például a “CCC” és “ATT” esetén, ahol, míg az egyik csoportban minimális a K-mer mennyisége, a másikban kimagasló.

Annak érdekében, hogy meghatározzuk a kialakult csoportok pontos számát, illetve azonosítsuk tagjait, a Pearson – korreláció eredményén hierarchikus klaszterelemzést végeztünk. Mivel a korábbi eredmények alapján a 4-es K-mernél tűntek legegységesebbek a csoporthatárok, ezt az értéket használtuk a klaszterek meghatározásakor is. Ahogy a Pearson - korrelációs mátrix (8. ábra) alapján is látszik, két nagy csoport volt elkülöníthető. Az egyik csoportba kerültek a 2-es és 6-os kromoszómán található HCR-ek, a másikba pedig az 1-es és 4-es kromoszómán lévők. A kisebb csoportok elemzése alapján az volt megfigyelhető, hogy az A és B HCR-ek mindegyik

kromoszómán közelebbi csoportokat alkotnak, míg a D jobban eltér a többitől a csoporton belül (10. ábra).



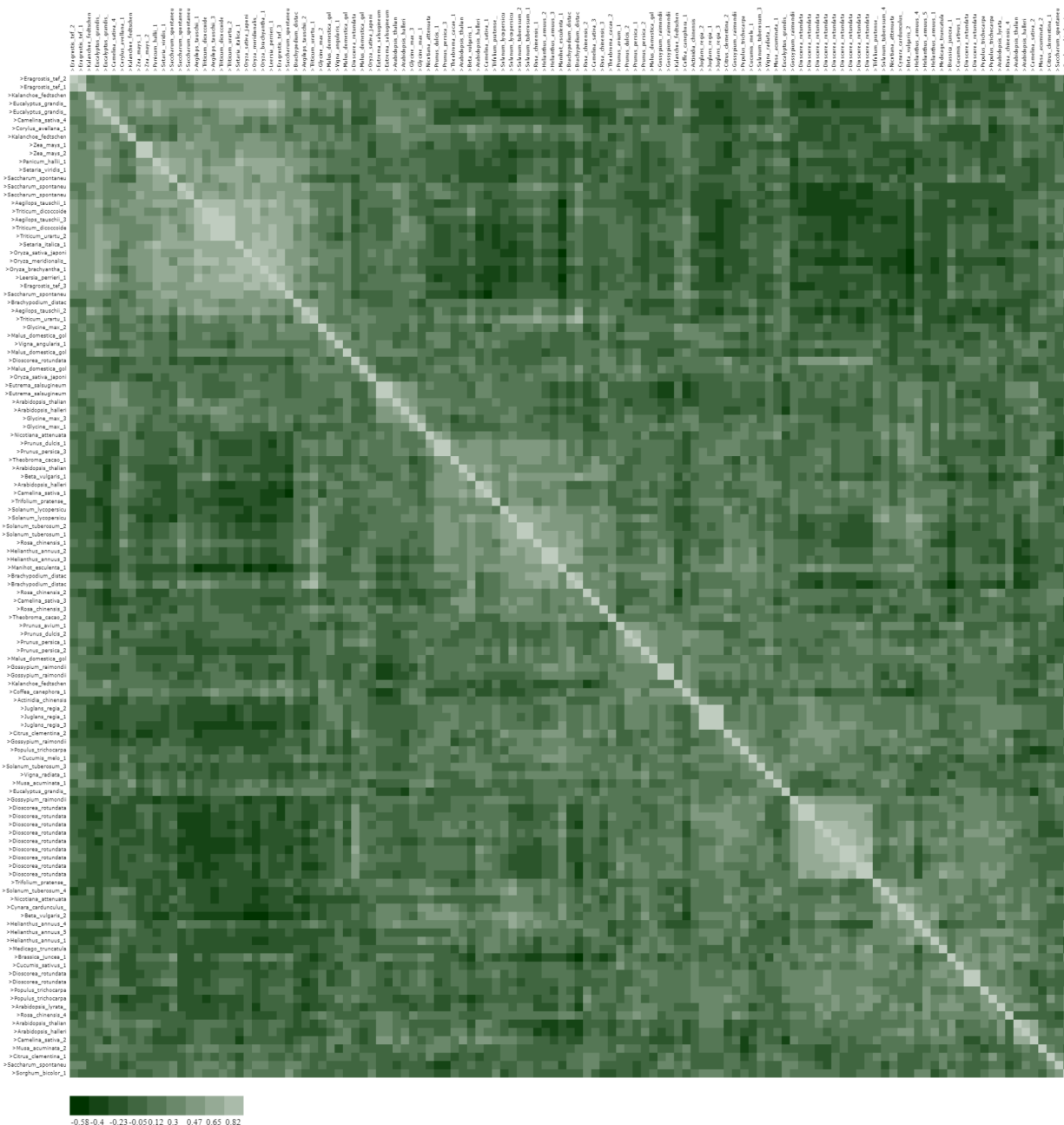
10. ábra: a *Triticum* fajok K-mer analízis alapján meghatározott klaszterei

A Pearson - korrelációs mátrix alapján elvégeztük a hierarchikus klaszterelemzést a *Triticum* fajok szekvenciáin. Ez alátámasztja a korábbi feltételezéseket, miszerint a szekvenciák a kromoszóma számuk alapján, nem pedig fajon belüli hasonlóságok alapján csoportosulnak. Megfigyelhető még emellett, hogy az egyes csoportokon belül több esetben is nagyobb hasonlóság fedezhető fel az A és B genom HCR-ei között, mint a D genoméval.

Fontos megjegyezni, hogy az allopoliploid fajok esetében a kromoszómaszerelvények a szülők ősi genomját hordozzák (*T. aestivum* két poliploidizációs esemény következtében alakult ki a *T. urartu* (AA), *Aegilops speltoides* (BB) és *T. tauschii* (DD) kereszteződéséből; az AABB hibrid 0,5 millió éve, az AABBDD pedig 10 ezer évvel ezelőtt jött létre); ez alatt az időintervallum alatt (különösen az első hibridizációs esemény óta) számos mutáció jöhetett létre (El Baidouri és mtsai., 2017). akár a GUT15 HCR régióján belül is, ami módosíthatta a sub-genom-specifikus HCR régiókat. Valóban ez figyelhető meg: az A és B genomok HCR homológjai jobban hasonlítanak egymásra, mint a D genom HCR tartalmú génje, éppen ahogy a két hibridizációs esemény alapján elvárt. Ez a megfigyelés arra enged következtetni, hogy egy lassú HCR homológ “uniformizáció zajlik”, következésképpen egy típusú funkciója lehet az összes HCR tartalmú homológoknak.

5.2.3. A HCR régió kívüli szakaszok K-mer tartalmának vizsgálata

A nem-HCR régió vizsgálatát azzal a céllal végeztük el, hogy azonosítsunk esetlegesen jelenlévő *cisz* elemeket a HCR régiók környezetében, amik részt vehetnek szabályozásukban. A K-mer analízist elvégeztük mind az upstream, mind a downstream szakaszokkal. Az upstream szakaszok hossza 76 nukleotidtól 2532-ig terjedt és konzerváltságot csak a fajokon belül mutatott. Ez volt elmondható a downstream szakaszcól is, aminek hossza 7-7000 nukleotidig változatos hosszúságú volt. Nem tudtunk jelentősen dúsuló K-mer mintázatot meghatározni (11. ábra).



11. ábra: a cDNS és ncRNS szekvenciák HCR előtti szakaszának Pearson - korrelációs mátrixa

A cDNS és ncRNS szekvenciák K-mer analízise során nem azonosítottunk jellegzetes, egymástól erősen eltérő K-mer ujjlenyomattal rendelkező csoportokat. A vizsgált szekvenciák a HCR szakaszon kívül heterogének, ezért vélhetően nem fehérjepartnereken keresztül vesznek részt a funkcióban

Eredményeink alapján feltételezhető, hogy a GUT15/CR20 HCR régióin kívül eső RNS nem vesz részt partner interakcióban vagy nem-specifikus partner kötésen keresztül működhet. Hipotézisünk, hogy ezek az RNS szegmensek vagy magának az lncRNS-ek a szabályozását látják

el (pl. burkolják a HCR-t, lokalizációt befolyásolnak, féléletidőt befolyásolnak stb.) vagy promiszkus kötések által szabályozzák a funkciót (pl. a foszfor-cukor lánc negatív töltése pozitív felületekkel rendelkező (fehérje)partnereket köt meg).

5.2.4. A cDNS/ncRNS szekvenciák K-mer analízise, hierarchikus klaszterelemzése és a kialakult csoportok vizsgálata

A DNS szekvenciák HCR régiójának elemzése megerősítette a megfigyelésünket, miszerint a vizsgált szekvenciák egyértelműen elkülöníthető hasonlósági csoportokat alkotnak, tehát a különböző csoportok eltérő fehérjepartnerekkel rendelkezhetnek. Bár a nagyszámú DNS minták használata során a K-mer mintázat jobban felismerhető, nem hanyagolható el annak a lehetősége, miszerint nem az összes DNS szekvenciában azonosított HCR-tartalmú gén fejeződik ki a transzkripció során (feltételezésünk szerint a GUT15/CR20 génnek aktív eleme az RNS maga, nem pedig a DNS). Azért, hogy ezekből az esetleges eltérő kifejeződésekből adódó hamis következtetéseket elkerüljünk, elvégeztük mind a K-mer analízist, mind a hierarchikus klaszterelemzést a cDNS/ncRNS szekvenciák HCR szakaszain. Ennek eredménye alátámasztotta a korábbiakat, a hasonló mintázatok, bár a kevesebb vizsgált szekvencia miatt kisebb mértékben, de hasonlóan jelentek bennük.

A cDNS/ncRNS szekvenciák vizsgálatának következő lépéseként meghatároztuk mind csak a HCR régiók, mind a teljes cDNS/ncRNS-ek K-mer ujjlenyomatait és az ezek alapján kialakult hasonlósági csoportokat elemeztük tovább. Korábbi tapasztalataink alapján ezeket a vizsgálatokat is 4-es K-mer értékkel végeztük el.

A HCR régiók elemzése alapján 8 hasonlósági csoportot különítettünk el. Ezeken újból elvégeztük a K-mer analízist, hogy megállapítsuk milyen K-merek jellemzőek a csoportokban. Az egyes csoportok összetétele rendkívül változatos volt, tagjaikat jellemzően nem a közeli rokon fajok alkották, a nagyobb kópiaszámmal rendelkező fajok pedig több csoportra oszlottak. Ezek az eredmények a korábbiakat támasztják alá, a K-mer ujjlenyomatok hasonlósága nem a rokonságokon alapul. Az egyes csoportok K-mer analízise azonban nem határozott meg olyan K-mereket, amik a csoport összes tagjában a többinél nagyobb gyakorisággal fordultak volna elő.

Hasonló eredményeket tapasztaltunk a teljes cDNS/ncRNS szekvenciák vizsgálatakor, azonban – mivel a teljes szekvenciák összességében jóval heterogénebbek voltak – 15 csoportot különítettünk el.

Összesítve tehát, a HCR régió és közvetlen környezetében magas szekvencia variabilitású K-mer mintázatokat találtunk, ami azt sugallja, hogy a HCR régió szerkezete, nem pedig a szekvencia-mintázat (K-mer mintázat) lehet a funkcionális tulajdonság.

5.2.5. *Arabidopsis thaliana* ncRNS szekvenciák K-mer analízise, hierarchikus klaszterelemzése és a kialakult csoportok vizsgálata

A következő lépésben az ismert *Arabidopsis thaliana* ncRNS-eket vizsgáltuk a korábbiakhoz hasonlóan K-mer analízissel és utána hierarchikus klaszterelemzéssel. Ezzel az volt a célunk, hogy meghatározzuk az *Arabidopsis thaliana* GUT15/CR20 lncRNS-eihez leginkább hasonló lncRNS-eket és az azokról a PlantsEnsemble adatbázisában elérhető információk alapján azonosítsunk potenciális fehérje partnereket. Ezeket a vizsgálatokat 5655 szekvencián végeztük el 3, 4, 5 és 6 K-mer méretekkel.

Mivel az általunk vizsgált lncRNS-ekhez leghasonlóbb ncRNS-eket szeretnénk volna meghatározni, 560 klasztert hoztunk létre az eredeti szekvenciákból, így kb a 10 leghasonlóbb ncRNS-t kaptuk meg (a klaszterelemzés 560 csoportot alakít ki, és mindegyik mintát a leghasonlóbb csoportba teszi, így előfordulhat, hogy nem mindegyik csoportba pontosan 10 ncRNS kerül). Érdekes módon, míg 3-as és 4-es K-mer értéknél a 3 lncRNS (GUT15, GUT15 splice variáns, CR20) a GUT15 splice variáns vált külön a másik kettőtől, 5-ös és 6-os K-mer érték esetén a CR20 került másik klaszterbe. Az ezzel a módszerrel meghatározott ncRNS-ek között azonosítottunk snoRNS-eket, illetve lncRNS-eket, azonban nem találtunk olyat, aminek a funkciója és/vagy fehérjepartnerei ismertek, lennének.

K-mer = 3		K-mer = 4		K-mer = 5		K-mer = 6	
AT1G08103.1	AT1G11175.1	AT1G05913.1	AT1G09407.1	AT1G03746.1	AT2G18440.1	AT1G04787.1	AT2G18440.1
AT2G18440.1	AT2G13665.1	AT1G06697.1	AT2G07885.1	AT1G26233.1	AT2G18440.2	AT1G07193.1	AT2G18440.2
AT2G22821.1	AT2G13665.2	AT1G06997.1	AT2G18440.2	AT1G74456.1	AT2G24592.1	AT1G07903.1	
AT3G02995.1	AT2G13665.3	AT1G09227.1	AT3G08340.1	AT2G04405.1	AT5G05705.1	AT1G08113.1	
AT4G05295.1	AT2G18440.2	AT2G06955.1	AT5G05515.1	AT3G03435.1		AT1G08303	
AT4G36648.1	AT4G07250.1	AT2G18440.1	AT5G06705.1	AT3G03445.1		AT1G09723.1	
AT5G01775.1	AT5G06705.1	AT3G04775.1	AT5G09635.1	AT3G09105.1		AT1G26233.1	
		AT3G07935.1		AT3G13525.1		AT1G60073	
		AT3G24614.1		AT3G58193.1		AT2G04405.1	
		AT4G05255.1		AT3G58196.1		AT2G04905.1	
		AT4G05315.1		AT4G07205.1		AT2G07405.1	
		AT4G14548.1		AT4G36648.1		AT2G08040.1	
		AT4G36648.1		AT5G01775.1		AT2G22496	
		AT5G02055.1		AT5G06565.1		AT3G02705.1	
		AT5G02645.1		AT5G08865.1		AT3G03665.1	
						AT3G07135.1	
						AT4G04665.1	
						AT4G05255.1	
						AT4G07175.1	
						AT4G36648.1	
						AT5G06965.1	
						AT5G08865.1	

1. táblázat: az *Arabidopsis thaliana* ncRNS-ein végzett hierarchikus klaszteranalízis eredményeinek összefoglalása

Az *Arabidopsis thaliana* kb 5600 ncRNS-én végeztünk K-mer analízist és annak eredményeiből hierarchikus klaszteranalízist, hogy azonosítsuk a GUT15 (at2g18440.1, zöld), GUT15 splice variáns (at2g18440.2, kék) és CR20 (at4g36648.1, rózsaszín) HCR-ekhez leghasonlóbb ncRNS-eket. K-mer 3-6 értékekkel végeztük el a vizsgálatot, mindegyik esetben 2 csoportba került a 3 HCR. Míg 3-as K-mer érték esetén a két csoport egyenlő tagszámú volt és GUT15 és a CR20 egy klaszterbe kerültek, 4-es érték esetén az arányok kis mértékben eltolódtak és ez a tendencia folytatódott a 6-os K-mer értékig, ahol csak a GUT15 és annak splice variánsa kerültek egy klaszterbe.

5.3 sORF-ek azonosítása a cDNS és ncRNS szekvenciákban, sORF funkcionalitás lehetőségének tesztelése

Az lncRNS-ek definíció szerint nem kódolnak fehérjét, azonban számos olyan lncRNS létezik mely mikro-peptid kódolásra képes. Mivel a HCR régió önmagában egy konzervált szakasz, a HCR régiót átívelő rövid leolvasási keretek (short Open Reading Frame, sORF) is (legalábbis részben) konzerváltak lehetnek. Hogy kiderítsük van-e elméleti lehetősége annak, hogy a HCR

tartalmú gének funkcionális terméke egy HCR-peptid, *in silico* transzláltuk a HCR régiót átfedő sORF-ket (5'-3' irányban, Start kodon + sORF + Stop kodon tulajdonságok jelenlétét megkövetelve) és elemeztük ezek konzerváltságát a törzsfejlődés folyamán. Magas variabilitást találtunk a peptidek között, mely a hajtúket összekötő szakaszok, a hajtúk csúcsi részének és a középső hajtú (poliA-poliU) hosszúságának eltérései és alacsony konzerváltsága miatt lépnek fel. Az alacsony konzerváltságot az *Arabidopsis* és közeli rokonán, a repce (*Brassica napus*) GUT15/CR20 homológ génjein példázzuk (megjegyzés: a repcében 12 GUT15 homológ gén van, az egyszerűség kedvéért 1 homológot mutatunk be ebben az esettanulmányban). A HCR régióban mindössze egy olyan sORF található meg, mely 10 aminosavnál hosszabb, start és stop kodonnal rendelkezik; ennek a peptidnek a konzerváltsága mindössze 22%, mely az első hajtú nukleotid konzerváltságából egyenesen következik, de a továbbiakban nem fenntartott (12. ábra).

GUT15_ORF2:	MSGALAWQVKLILNKKKILWVLERRSHGLLFFPGISSPLCVSSCLCSISLP-ALSHFIS	59
CR20_ORF2:	MTGALAWQVTK-ILNKKKKFRVVEGSSHGFFLPGFSPACVSCFSLFVHTTFDLDSFT-	58
BnaGUT15-chrA8_ORF2:	MTGALAWQVTL-NNNNKNCFVLVRGVRTGDFSPGSPFTLCLALS-----	45
	*:*****. **: : : :. * : * * *: :	
GUT15_ORF2:	FLNAHIHSKTDHKQSL	75
CR20_ORF2:	-----	58
BnaGUT15-chrA8_ORF2:	-----	45

12. ábra: az AtGUT15, AtCR20 és BnaGUT15 (chrA8) gének HCCR régióival átfedő sORF peptidek hasonlósági vizsgálata (a csillag azonos aminosavat, a kettőspont hasonló aminosavat jelöl).

Ezen eredmények alapján azt gondoljuk, hogy a GUT15/CR20 lncRNS géncsalád funkcionális egysége nem mikro-peptid, hanem az lncRNS molekula.

6. Összefoglalás

Míg korábban úgy vélték, hogy a hosszú nem-kódoló RNS-ek (lncRNS-ek) csak transzkripcionális melléktermékek, mára egyre több lncRNS-eről bizonyosodik be, hogy szabályozott módon keletkezik és jól meghatározott funkciója van. Mivel azonban az lncRNS-ek vizsgálata konzerváltságuk hiánya miatt nagyon nehéz, egy olyan vizsgálati eszköztárra van szükség, mely alkalmas a funkcionális kandidánsok kiválasztására.

Munkánk során egy részben konzervált lncRNS családot vizsgálunk meg bioinformatikai eszköztárral.

Következtetések az alábbiak:

1. A GUT15/CR20 géncsaládban fellelhető HCR régió erőteljesen konzervált a törzsfejlődés során zárwatermőkben;
2. A GUT15/CR20 génekben megtalálható HCR régió egy konzervált másodlagos szerkezettel rendelkezik.
3. A HCR régió kívül eső szakasz nem tartalmaz dúsuló K-mer ujjlenyomatot, magas variabilitású és változó hosszúságú, a GUT15/CR20 lncRNS molekulán belül a HCR régió elhelyezkedése random; mindezek alapján feltételezzük, hogy a non-HCR régiók nem vesznek rész specifikus partner kölcsönhatásban, hanem vagy aspecifikus kötés hoznak létre (töltésük révén), vagy a HCR struktúra kifejeződésének, szabályozásának eszközei (éréshez szükségesek, féléteidőt szabályoznak, lokalizációt módosítanak stb.)
4. A HCR-tartalmú gének száma változott (legtöbb esetben nőtt) a törzsfejlődés folyamán, feltehetően többszörös, független események következtében; feltételezzük, hogy a GUT15/CR20 lncRNS hatásmechanizmusa dózisfüggő, vagy több gén/traszkriptum közreműködésén alapul.
5. A GUT15/CR20 lncRNS-ek nem kódolnak konzervált sORF-et.

Összegezve, a GUT15/CR20 gének biológiai funkciója nagy valószínűséggel a HCR régióhoz kötődik és annak struktúrája révén valósul meg. A HCR régió magas fokú konzerváltsága arra enged következtetni, hogy a zárwatermők őseiben jelent meg egy alkalommal és konzerválódott, és a zárwatermő életmódhoz kötött ősi funkciót lát el.

Ezek fényében a jövőben kísérletet fogunk tenni a meghatározott csoportok HCR-(fehérje)partnereinek felkutatására és pontos funkciójuk meghatározására, ezzel elősegítve a magas hőmérsékleti stressz során betöltött funkciójuk megértését.

A munkafolyamat melyet a munkánkban alkalmaztunk, a jövőben alkalmas lehet más lncRNS-ek vizsgálatára is.

7. Köszönetnyilvánítás

Köszönettel tartozom témavezetőmnek, Dr. Csorba Tibornak, valamint Szaker Henriknek a szakmai iránymutatásért és a közös munkáért.

8. Irodalomjegyzék

- Ambro van Hoof, James P, Kastenmayer CBT and PJG. GUT15 cDNAs from Tobacco (Accession No. U84972) and Arabidopsis (Accession No. U84973) Correspond to Transcripts with Unusual Metabolism and a Short Conserved ORF. *Plant Physiol.* 1997;113:1004.
- Ariel, F., Jegu, T., Latrasse, D., Romero-Barrios, N., Christ, A., Benhamed, M., & Crespi, M. (2014). Noncoding transcription by alternative rna polymerases dynamically regulates an auxin-driven chromatin loop. *Molecular Cell*, 55(3), 383–396. <https://doi.org/10.1016/j.molcel.2014.06.011>
- Ariel, F., Lucero, L., Christ, A., Mammarella, M. F., Jegu, T., Veluchamy, A., Mariappan, K., Latrasse, D., Blein, T., Liu, C., Benhamed, M., & Crespi, M. (2020). R-Loop Mediated trans Action of the APOLO Long Noncoding RNA. *Molecular Cell*, 77(5), 1055-1065.e4. <https://doi.org/10.1016/j.molcel.2019.12.015>
- Bussotti, G., Notredame, C., & Enright, A. J. (2013). Detecting and comparing non-coding RNAs in the high-throughput era. In *International Journal of Molecular Sciences* (Köt. 14, Szám 8, o. 15423–15458). <https://doi.org/10.3390/ijms140815423>
- Carlevaro-Fita, J., Rahim, A., Guigó, R., Vardy, L. A., & Johnson, R. (2016). Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA*, 22(6), 867–882. <https://doi.org/10.1261/rna.053561.115>
- Cech, T. R., & Steitz, J. A. (2014). The noncoding RNA revolution - Trashing old rules to forge new ones. *Cell*, 157(1), 77–94. <https://doi.org/10.1016/j.cell.2014.03.008>
- Choi, S. W., Kim, H. W., & Nam, J. W. (2019). The small peptide world in long noncoding RNAs. In *Briefings in Bioinformatics* (Köt. 20, Szám 5, o. 1853–1864). Oxford University Press. <https://doi.org/10.1093/bib/bby055>
- El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., Alaux, M., Quesneville, H., Pont, C., & Salse, J. (2017). Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytologist*, 213(3), 1477–1486. <https://doi.org/10.1111/nph.14113>
- Fallmann, J., Will, S., Engelhardt, J., Grüning, B., Backofen, R., & Stadler, P. F. (2017). Recent advances in RNA folding. In *Journal of Biotechnology* (Köt. 261, o. 97–104). Elsevier B.V. <https://doi.org/10.1016/j.jbiotec.2017.07.007>
- Fedak, H., Palusinska, M., Krzyczmonik, K., Brzezniak, L., Yatusевич, R., Pietras, Z., Kaczanowski, S., & Swiezewski, S. (2016). Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48), E7846–E7855. <https://doi.org/10.1073/pnas.1608827113>

- Henriques, R., Wang, H., Liu, J., Boix, M., Huang, L. F., & Chua, N. H. (2017). The antiphasic regulatory module comprising CDF5 and its antisense RNA FLORE links the circadian clock to photoperiodic flowering. *New Phytologist*, 216(3), 854–867. <https://doi.org/10.1111/nph.14703>
- Heo, J. B., & Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science*, 331(6013), 76–79. <https://doi.org/10.1126/science.1197349>
- Hickson, R. E., Simon, C., & Perrey, S. W. (2000). The Performance of Several Multiple-Sequence Alignment Programs in Relation to Secondary-Structure Features for an rRNA Sequence. *Mol. Biol. Evol.*, 17(4), 530–539. <https://academic.oup.com/mbe/article/17/4/530/1127647>
- Horner, D. S., & Pesole, G. (2004). Phylogenetic analyses: A brief introduction to methods and their application. In *Expert Review of Molecular Diagnostics* (Köt. 4, Szám 3, o. 339–350). <https://doi.org/10.1586/14737159.4.3.339>
- Hsu, P. Y., Calviello, L., Wu, H.-Y. L., Li, F.-W., Rothfels, C. J., Ohler, U., & Benfey, P. N. (é. n.). *Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis*. <https://doi.org/10.5061/dryad.m8jr7>
- Jin, Y., Ivanov, M., Dittrich, A. N., Nelson, A. D., & Marquardt, S. (2023). <scp>LncRNA FLAIL</scp> affects alternative splicing and represses flowering in *Arabidopsis*. *The EMBO Journal*. <https://doi.org/10.15252/embj.2022110921>
- Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J., & Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, 50(10), 1474–1482. <https://doi.org/10.1038/s41588-018-0207-8>
- Kowalczyk, J., Palusinska, M., Wroblewska-Swiniarska, A., Pietras, Z., Szewc, L., Dolata, J., Jarmolowski, A., & Swiezewski, S. (2017). Alternative Polyadenylation of the Sense Transcript Controls Antisense Transcription of DELAY OF GERMINATION 1 in *Arabidopsis*. In *Molecular Plant* (Köt. 10, Szám 10, o. 1349–1352). Cell Press. <https://doi.org/10.1016/j.molp.2017.07.011>
- Kruszka, K., Pieczynski, M., Windels, D., Bielewicz, D., Jarmolowski, A., Szweykowska-Kulinska, Z., & Vazquez, F. (2012). Role of microRNAs and other sRNAs of plants in their changing environments. *Journal of Plant Physiology*, 169(16), 1664–1672. <https://doi.org/10.1016/j.jplph.2012.03.009>
- Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L. L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., ... Wu, M. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*. <https://doi.org/10.1038/s41580-022-00566-8>

- Moison, M., Pacheco, J. M., Lucero, L., Fonouni-Farde, C., Rodríguez-Melo, J., Mansilla, N., Christ, A., Bazin, J., Benhamed, M., Ibañez, F., Crespi, M., Estevez, J. M., & Ariel, F. (2021). The lncRNA APOLO interacts with the transcription factor WRKY42 to trigger root hair cell expansion in response to cold. *Molecular Plant*, 14(6), 937–948. <https://doi.org/10.1016/j.molp.2021.03.008>
- Morris, K. V., & Mattick, J. S. (2014). The rise of regulatory RNA. In *Nature Reviews Genetics* (Köt. 15, Szám 6, o. 423–437). Nature Publishing Group. <https://doi.org/10.1038/nrg3722>
- Peschansky, V. J., & Wahlestedt, C. (2014). Non-coding RNAs as direct and indirect modulators of epigenetic regulation. In *Epigenetics* (Köt. 9, Szám 1, o. 3–12). Taylor and Francis Inc. <https://doi.org/10.4161/epi.27473>
- Phillips, A., Janies, D., & Wheeler, W. (2000). Multiple sequence alignment in phylogenetic analysis. In *Molecular Phylogenetics and Evolution* (Köt. 16, Szám 3, o. 317–330). <https://doi.org/10.1006/mpev.2000.0785>
- Plewka, P., Thompson, A., Szymanski, M., Nuc, P., Knop, K., Rasinska, A., Bialkowska, A., Szweykowska-Kulinska, Z., Karlowski, W. M., & Jarmolowski, A. (2018). A stable tRNA-like molecule is generated from the long noncoding RNA GUT15 in Arabidopsis. *RNA Biology*, 15(6), 726–738. <https://doi.org/10.1080/15476286.2018.1445404>
- Rai, M. I., Alam, M., Lightfoot, D. A., Gurha, P., & Afzal, A. J. (2019). Classification and experimental identification of plant long non-coding RNAs. *Genomics*, 111(5), 997–1005. <https://doi.org/10.1016/j.ygeno.2018.04.014>
- Shih, J. W., & Kung, H. J. (2017). Long non-coding RNA and tumor hypoxia: New players ushered toward an old arena. In *Journal of Biomedical Science* (Köt. 24, Szám 1). BioMed Central Ltd. <https://doi.org/10.1186/s12929-017-0358-4>
- Sunkar, R., Li, Y., & Jagadeeswaran, G. (2012). Functions of microRNAs in plant stress responses. *Trends in Plant Science*, 17(4), 196–203. <https://doi.org/10.1016/j.tplants.2012.01.010>
- Swiezewski, S., Liu, F., Magusin, A., & Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*, 462(7274), 799–802. <https://doi.org/10.1038/nature08618>
- Taylor, C. B., & Green, P. J. (1995). Identification and characterization of genes with unstable transcripts (GUTs) in tobacco. In *Plant Molecular Biology* (Köt. 28).
- Teramoto, H., Toyama, T., Takeba, G., & Tsuji, H. (1995). Changes in expression of two cytokinin-repressed genes, CR9 and CR20, in relation to aging, greening and wounding in cucumber. In *Planta*. Springer-Verlag.
- Teramoto, H., Toyama, T., & Tsuji, H. (1996). Noncoding RNA for CR20, a cytokinin-repressed gene of cucumber. In *Plant Molecular Biology* (Köt. 32). KluwerAcademic Publishers.

- Wang, J., Meng, X., Dobrovolskaya, O. B., Orlov, Y. L., & Chen, M. (2017). Non-coding RNAs and Their Roles in Stress Response in Plants Wang J et al / miRNA and lncRNA in Plant Stress Response. In *Genomics, Proteomics and Bioinformatics* (Köt. 15, Szám 5, o. 301–312). Beijing Genomics Institute. <https://doi.org/10.1016/j.gpb.2017.01.007>
- Wunderlich, M., Groß-Hardt, R., & Schöffl, F. (2014). Heat shock factor HSFB2a involved in gametophyte development of *Arabidopsis thaliana* and its expression is controlled by a heat-inducible long non-coding antisense RNA. *Plant Molecular Biology*, 85(6), 541–550. <https://doi.org/10.1007/s11103-014-0202-0>
- Ylstra, B., & McCormick, S. (1999). Analysis of mRNA stabilities during pollen development and in BY2 cells. *Plant Journal*, 20(1), 101–108. <https://doi.org/10.1046/j.1365-313X.1999.00580.x>