DIPLOMADOLGOZAT

GÁL NÓRA Mezőgazdasági biotechnológus Msc

Gödöllő 2023

Magyar Agrár- és Élettudományi Egyetem Szent István Campus Mezőgazdasági biotechnológus Szak

A hőindukált GUT15/CR20 lncRNS család vizsgálata bioinformatikai módszerekkel

Belső konzulens:	Dr. Csorba Tibor Levente		
	csoportvezető,		
	tudományos főmunkatárs		
Készítette:	Gál Nóra		
	JKUW77		
	levelező tagozat		
Intézet/Tanszék:	Genetika és Biotechnológia Intézet, Növénybiotechnológia tanszék		

Gödöllő 2023

Tartalomjegyzék

1. Bevezetés, célkitűzések
2. Irodalmi áttekintés
2.1. A nem-kódoló RNS-ek 5
2.1.1 A hosszú nem-kódoló RNS-ek
2.1.2. GUT15/CR20 lncRNS család 10
2.2 lncRNS-ek vizsgálati lehetőségei 11
2.2.1. Szekvenciaillesztés
2.2.2. Filogenetikai elemzés
2.2.3. K-mer analízis
3. Anyag és módszer
3.1 Számítógépes környezet
3.2. Szekvenciaillesztés
3.3. Filogenetikai elemzés
3.4. Másodlagos szerkezet meghatározása16
3.5. K-mer analízis
3.6. Hierarchikus klaszterelemzés 17
3.7. ORF-ek azonosítása 17
4. Eredmények és megvitatásuk
4.1. Szekvenciaillesztés és filogenetikai elemzés 19
4.1.1 Szekvenciaillesztés 19
4.1.2 Filogenetikai fa készítése
4.1.3. A HCR régió szerkezeti elemzése
4.2. K-mer analízis, hierarchikus klaszterelemzés
4.2.1. A HCR régió K-mer tartalmának vizsgálata
4.2.2. Triticum fajok K-mer analízise és hierarchikus klaszterelemzése
4.2.3. A HCR régión kívüli szakaszok K-mer tartalmának vizsgálata
4.2.4. A cDNS/ncRNS szekvenciák K-mer analízise, hieararchikus klaszterelemzése és a kialakult csoportok vizsgálata

4.2.5. Arabidopsis thaliana ncRNS szekvenciák K-mer analízise, hierarchikus klaszterelemzése és a kialakult csoportok vizsgálata	40
4.3 sORF-ek azonosítása a cDNS és ncRNS szekvenciákban, sORF funkcionalitás lehetőségének tesztelése	41
5. Következtetések és javaslatok	
6. Összefoglalás	
7. Köszönetnyilvánítás	
8. Irodalomjegyzék	47

1. Bevezetés, célkitűzések

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 fehérjét nem kódol, RNS-ként funkcionálhat. Az 1950-es években és ezt követően, számos fehérjét nem kódoló RNS (non-coding RNA, 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 (small RNA, sRNA) 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, hormonkoncentráció), vagy külső környezeti tényezők (pl. biotikus vagy abiotikus stressz) hatására. Emellett sok lncRNS érési folyamatokon (pl. intron kivágódás) is keresztülmegy. 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 (transzkripciós melléktermék), kialakulóban lévő "*szemi*"-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.

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 lókusz (*At1g19397*, 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, *cisz* elemek) azonosítására és ezáltal a feltételezhető funkció meghatározására;

(iii) az lncRNS fehérjét kódoló homológjainak az azonosításával és összehasonlításával feltérképezzük a lehetséges evolúciós utat, mely a GUT15/CR20 lncRNS család kialakulásához vezethetett (*lásd később*, 2. ábra);

(iv) mindezeken túlmenően, egy olyan automatizált bioinformatikai munkafolyamatot ("workflow"-t) szeretnénk létrehozni, amely alkalmazásával átláthatóbb és könnyebben vizsgálható más lncRNS-ek konzerváltsága, funkcionalitása és evolúciója is.

2. Irodalmi áttekintés

2.1. A nem-kódoló RNS-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 (J. 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 expressziója folyamatos, szerepe sok esetben jól ismert (Romano és mtsai., 2017). Legismertebb képviselőik 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 (Ponting és mtsai., 2009).

A szabályozó funkciót ellátó ncRNS-ek funckiója a háztartási RNS-ekkel szemben kevésbé ismert, vizsgálatuk azonban az elmúlt években egyre nagyobb hangsúlyt kapott (Romano és mtsai., 2017). A szabályozó ncRNS-ek csoportosítása méret szerint történik, ebből adódóan heterogén funkciókkal rendelkeznek. Megkülönböztetünk kis RNS-eket (sRNS) és hosszú nem-kódoló RNS-eket (lncRNS)(Morris & Mattick, 2014)(1. ábra). Az sRNS-ek jellemzően 21-24 nukleotid (nt) hosszúságúak és számos funkcióval rendelkezhetnek, részt vesznek többek között az RNS-interferencia folyamatában és a tápanyag homeosztázis kialakításában is (Kruszka és mtsai., 2012; Pattanayak és mtsai., 2013), emellett egyre több tanulmány igazolja szerepüket a növényi stresszválasz kialakításában is, például a miR393, miR160 és miR167 sRNS-ek fokozott expressziója szárazság és/vagy sóstressz esetén számos növényfajban nyert igazolást (Sunkar és mtsai., 2012).

2.1.1 A hosszú nem-kódoló RNS-ek

A nem-kódoló RNS átiratok legnagyobb csoportját a hosszú nem-kódoló RNS-ek (lncRNS) képviselik (Mattick & Makunin, 2006). Ezek pontos meghatározása azonban rendkívüli sokféleségük miatt nehézkes. Biogenezisük és szerkezetük alapján sok hasonlóságot mutatnak az mRNS-ekkel (pl PolII általi transzkripció, poli-adenil farok), azonban számos lncRNS esetén jelentős eltérések is (pl eltérő 3'-terminális feldolgozás, nyílt olvasókeretek (open reading frame, ORF) hiánya) felfedezhetők. Ezen eltérések megléte és mértéke azonban nagymértékű divergenciát mutat az lncRNS-eken belül. Összességében a legelfogadottabb meghatározás alapján lncRNS-nek az olyan, legalább 200 nukleotid hosszúságú transzkriptomokat nevezzük, amik nem kódolnak fehérjét (Mattick és mtsai., 2023; Quinn & Chang, 2016a).

Míg az mRNS-ek meghatároznak egy fehérjét, a fehérje pedig rendelkezik egy funkcióval, az lncRNS-ek maguk alkotják a funkcionális egységet. Az lncRNS-ek az mRNS-ekéhez képest alacsonyabb expressziós szinttel rendelkeznek, átírásukat azonban hozzájuk hasonlóan a PolII végzi, de a növény specifikus Pol IV és Pol V is termelhetnek lncRNS-eket (Li és mtsai., 2014; Wierzbicki és mtsai., 2008).

Mivel az lncRNS-ek az mRNS-ekéhez képest alacsonyabb expressziós szinttel rendelkeznek, szekvenciájuk pedig a legtöbb esetben a közeli fajokon kívül nem mutat konzerváltságot, fontosságukat sokáig megkérdőjelezték. Egyre több esetben bizonyosodott be azonban, hogy kifejeződésük szövetspecifikus, illetve változni képes a külső, környezeti tényezők (pl. stressz) hatására. Az lncRNS kifejeződésének specifikus mintázatai koordinálják a sejtek állapotát, differenciálódását, fejlődését és betegségeit (Batista & Chang, 2013; Flynn & Chang, 2014). Az lncRNS-ek többségének funkciói ismeretlenek, soknak talán nincs is értékelhető funkciója, de néhány szerepe és hatásmechanizmusa jól ismert, mint például az X inactive specific transcript (XIST), a HOX transzkript antisense RNS (HOTAIR) és a telomerase RNS component (TERC) lncRNS-eké és ezek listája folyamatosan bővül (Geisler & Coller, 2013).

Az lncRNS-eket a fehérjekódoló génekhez viszonyított átírási helyük alapján öt csoportba oszthatjuk. Ez alapján elkülöníthetőek a természetes antiszensz lncRNS-ek (natural antisense transcripts, NAT-lncRNAs), amik a fehérjekódoló génekkel ellentétes irányban íródnak át, az intron régióról származó (intronic ncRNA), az enhancer régióról származó (enRNAs), a promoter-asszociált transzkriptumok (promoter ncRNAs) és a hosszú intergénikus ncRNS-ek (long intergenic RNAs, lincRNAs)(Shih & Kung, 2017).

Korábbi kutatások alapján több, mint 200 *Arabidopsis thaliana* transzkriptom adat elemzése 40000 feltételezett lncRNS-t azonosított, amik között több mint 30000 NAT-ot és több mint 6000 lincRNS-t. Ezeknek az lncRNS-eknek a többsége nem társul sRNS-ekhez és transzkripciós szintjük, az emlősökhöz hasonlóan alacsonyabb az mRNS-ekénél. A NAT-párok, tehát kódoló (vagy nem kódoló) gének antiparalell száláról keletkező lncRNS-ek meglepően gyakoriak az *Arabidopsis*ban; fehérjekódoló lókuszainak 70%-a 200-12370 nt között potenciális NAT-párokat kódol. Az Arabidopsis NAT lncRNS-ek expressziója erősen szövetspecifikus és sok esetben reagál a biotikus, vagy abiotikus stresszhatásokra (J. Jin és mtsai., 2013; Liu és mtsai., 2012; H. Wang és mtsai., 2014).

Az lncRNS-ek hatása kifejeződhet a keletkezési helyükön, vagyis "*cisz*"-ben, vagy távoli lókuszokon azaz "*transz*"-ban. A *cisz-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 "mikropeptid" a funkcionális egység (Choi és mtsai., 2019). Az lncRNS-ek általában nem önmagukban, hanem több molekulából összeszerelődő komplexek alegységeiként működnek; lncRNS-ek partnerei lehetnek specifikus genomi lókuszok (DNS), lehetnek RNS-ek vagy fehérjék. Az lncRNS komplexek epigenetikai, transzkripcionális, poszt-transzkripcionális, transzlációs szinten szabályozhatnak számtalan folyamatot (Quinn & Chang, 2016b).

Az lncRNS-ek legismertebb funkciói a transzkripció szabályozójaként betöltött szerepük, például a PoIII szabályozásán keresztül, de több állati lncRNS-nél is megfigyelték, hogy elősegítik a transzkripciós faktorok (TF-ek) foszforilációját, és így szabályozzák DNS-kötő aktivitásukat (Pin Wang és mtsai., 2014). Számos eukarióta lncRNS a transzkripció iniciációjának és elongációjának szabályozásán keresztül fejti ki hatását, illetve sok esetben befolyásolhatják a kromatin topológiáját és a nukleáris szerveződést is (Bonasio & Shiekhattar, 2014). Például az *Arabidopsis* transz-hatású lncRNS HID1 társul a TF gén PIF3 kromatinjához és gátolja annak átíródását (Y. Wang és mtsai., 2014).

Az lncRNS-ek evolúciójának magyarázatára számos mechanizmust javasoltak, köztük a de novo keletkezést, a duplikációt és a fehérjekódoló génekből vagy transzpozíciós elemekből (TE-k) való exaptációt (Hezroni és mtsai., 2015). Ezeket a mechanizmusokat különböző tanulmányok megerősítették, amelyek új lncRNS-eket azonosítottak különböző fajokban (Derrien és mtsai., 2012), és bizonyítékot szolgáltattak lncRNS-duplikációs eseményekre (Marques & Ponting, 2009;

Ulitsky és mtsai., 2011) és TE-kből való exaptációra (Kelley & Rinn, 2012; Sun és mtsai., 2013). Összehasonlító genomikai elemzések kimutatták, hogy az lncRNS-ek evolúcióját alacsony szekvencia konzerváció, gyenge tisztító szelekció és az lncRNS-repertoárok gyors fluktuációja jellemzi (Pauli és mtsai., 2012; Ulitsky & Bartel, 2013). E megfigyelések ellenére egyes lncRNSek erős szekvencia-konzerváltságot és funkcionális korlátozást mutatnak. Például az Xkromoszóma inaktiválásában (Xist) és a genomiális imprintingben (H19) részt vevő lncRNS-ekről kimutatták, hogy erős tisztító szelekció alatt állnak (Gabory és mtsai., 2010; Laurent Durent és mtsai., 2006). Ezen túlmenően a fajok közötti funkcionális konzerváció vizsgálata kimutatta, hogy a konzervált másodlagos struktúrájú lncRNS-ek általában konzervált funkciókkal rendelkeznek, míg konzervált szekvenciájúak nem feltétlenül (Quinn & Chang, 2016a). a Az lncRNS-ek evolúcióját nagyfokú szövet- és fajspecifikusság is jellemzi (Mercer és mtsai., 2008). Ez a specifitás származhat új funkciók vagy szövetspecifikus szerepek megszerzéséből vagy az ősi funkciók elvesztéséből. Ezenkívül az lncRNS-ek szövetspecifikus kifejeződését gyakran specifikus transzkripciós faktorok, kromatin-módosítás és más epigenetikai mechanizmusok szabályozzák (Rinn & Chang, 2012). Ezek a mechanizmusok szintén szerepet játszhatnak az lncRNS-ek evolúciójában, mivel elősegítik az új funkcionális lncRNS-ek kialakulását a nem funkcionális szekvenciákból.



1. ábra: A nemkódoló RNS-ek csoportosítása

Az ncRNS-ek fehérjét nem kódoló átiratok. 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 sejtfunkciók 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 (intronic 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 organizmusok**ban (*Arabidopsis thaliana*, repce, paradicsom, árpa, búza stb.) is. Legtöbbnek a 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 lókuszának 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 (Y. 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 szabályozásában (Wunderlich és mtsai., 2014).

2.1.2. GUT15/CR20 lncRNS család

A zárvatermőkben 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 et al., 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 szövetspecifikus, pollen sejtekben markánsan megemelkedik (Ylstra & McCormick, 1999). A valós szerep megállapítása érdekében a *Green labor gut15;cr20* dupla mutáns *Arabidopsis* növényeket vizsgáltak, és ABA hiper-szenzitívnek találták őket magas hőmérsékleti körülmények között (ABA+30°C), ebből adódóan környezeti változáshoz való alkalmazkodási feladatot feltételeztek ezeknek az lncRNS-eknek (Lu és mtsai., 2003).

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 géncsaládról (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.

2.2 IncRNS-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 nehézkesebb. Az lncRNS génekben nem fellelhető a *start* kodon (ATG), a *stop* kodon (TGA, TAG vagy TAA), nincs bennük *nyílt leolvasási keret* (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 fejleszetettek ki.

2.2.1. Szekvenciaillesztés

A szekvenciaillesztés, avagy BLAST (Basic local alignment search tool) egy szekvencia hasonlóság kereső program, amely a felhasználó által megadott szekvencia egy szekvenciaadatbázissal való összehasonlítására használható (Altschup és mtsai., 1990). A BLAST rövid egyezéseket talál két szekvencia között, és ezekből a "hot spot"-okból képzi a szekvenciaillesztéseket. A BLAST az illesztések elvégzése mellett statisztikai információt is szolgáltat az illesztésekről; ez az "expect" (várható) érték ("*e* érték") vagy a false-pozitív arány (Jian Ye és mtsai., 2006).

2.2.2. 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), de ugyanakkor olyan egyedi jellemzőkkel 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 ős-szekvenciából 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 ("lötyögő nukleotid") megváltozása sok esetben nem okoz aminosav változást, míg a tripleten 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).

2.2.3. 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 megbízható információval. Mindemellett, a homológ lncRNS-ek funkciójának ismerete (más organizmusban) sem elég önmagában a vizsgált lncRNS funkciójának megértéséhez.

Ezek következtében az IncRNS-ek vizsgálata eltérő technikákat igényel. Mivel az IncRNS-ek funkcióját feltehetőleg a hozzájuk kapcsolódó RNS-kötő fehérje partnerek határozzák meg, ezek pedig rövid, 3-8 bp hosszú szakaszokhoz (*cisz* 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 IncRNS-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).

3. Anyag és módszer

3.1 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 a Python 3.8 verziót használtuk hozzá.

3.2. Szekvenciaillesztés

A lokális szekvenciaillesztést az Ensembl Plants (Ensembl Plants) adatbázisában található genomi, cDNS és ncRNS szekvenciákon az AtGUT15 (AT2G18440) konzervált régiójának első 37 bázis hosszú szekvenciája alapján végeztük (GACCTTTGCCATGTCAGGTGCGCTTGCATGGC AGGTC). A szekvenciaillesztésez az NCBI Blast+ alkalmazást használtuk 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 vissza, 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, kb 50 nt GT-gazdag régiót, illetve 10 nt downstream szakaszt (3. ábra). 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.

15

3.3. 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 eszközzel (<u>MAFFT alignment and NJ / UPGMA phylogeny</u> (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 (<u>Common Taxonomy Tree (nih.gov</u>)).

Az eredményeket mindkét esetben a MAFFT által ajánlott Phylo.io-val ábrázoltuk, majd newick (.nw) fájlként az R ggplot2, ggtree, cowplot, seqinr, ade4 és ape csomagok használatával hasonlítottuk össze.

3.4. Másodlagos szerkezet meghatározása

Az lncRNS-ek másodlagos és harmadlagos szerkezetének meghatározását a cDNS/ncRNS mintákon, a legszűkebb értelemben vett, 120 bázis hosszú HCR régión 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.

3.5. K-mer analízis

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

3.6. Hierarchikus klaszterelemzés

A K-mer analízis során elvégzett Pearson-korreláció eredményeit .csv fájlként felhasználva végeztük el a hierarchikus klaszterelemzést, az R "amap" csomag használatával és a "ggplot" és "phytools" csomagok segítségével ábrázoltuk.

3.7. 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.

3.8. Diagramok, ábrák létrehozása

Az ábrákat és diagramokat (a fent említetteken kívül) az R (4. ábra) és Python programnyelveken (11., 13. ábra), Excel táblázatkezelőben (1. táblázat, 5. ábra), illetve a Jalview (3. ábra) és az Inkscape (1., 2., 3. ábra) programokban készítettem el.



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.

4. Eredmények és megvitatásuk

4.1. 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-tól 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.

4.1.1 Szekvenciaillesztés

A szekvenciaillesztést az adatbázisokban elérhető összes faj szekvenciáin elvégeztük. Ez legfőképp zárvatermőket jelentett, ezeken kívül pedig alga és moha fajok adatait. Összesen 102 faj genomi (DNS adat), 105 faj transzkriptom (cDNS adat) és 57 faj ncRNS-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 kizárólag algák és mohák közé tartoztak (a fent említett 10 faj cDNS és ncRNS szekvenciáiban sem volt fellelhető a HCR szekvencia). Továbbá az adatbázisokban megtalálható nyitvatermő fajok szekvenciáiban sem azonosítható HCR szekvencia (DNS és cDNA/ncRNS 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. 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 régión kívüli szekvenciák határozhatják meg.

4.1.2 Filogenetikai fa készítése

Hogy rátekintést nyerjünk a HCR régió változására/evolúciójára a törzsfejlődés során, létrehoztunk egy hasonlóságon alapuló HCR filogenetikai fát és ezt összevetettük a növényi törzsfejlődési 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 ábrázoltuk. A filogenetikai fákat 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.A 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 rokonaik mellett helyezkednek el, a rendek pozíciója nem egyezik a taxonómia 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.A, B ábra).





6. ábra: a genomi (A) és cDNS/ncRNS (B) HCR szekvenciák alapján készített filogenetikai fák Az azonosított HCR-eknek fajonként meghatároztuk a konszenzus szekvenciáját, majd ez alapján többszörös szekvenciaillesztést végezve létrehoztuk a szekvenciák hasonlóságán alapuló filogenetikai fákat. A filogenetikai fákon különböző színekkel jelöltük az eltérő nemzetségekbe tartozó fajokat (lsd jelmagyarázat). A filogenetikai fák 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 is készítettünk egy taxonómiai fát (6.B ábra), amit összevetettünk az érintett fajok (tehát a cDNS szekvenciákban HCR-t tartalmazó fajok) törzsfejlődési 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 tapasztaltunk: az egyes taxonómiai csoportokba tartozó fajok (ahogy a DNS szekvenciák esetén is) egy/közeli 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 ősében 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 adhat 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 ilyet egy esetben sem találtuk).

4.1.3. 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 hajtű 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 hajtű a legkonzerváltabb, míg a második és harmadik kevésbé. A második hajtű duplaszálú régióját egy reverz-komplementer (többnyire) poli-A és egy poli-U nukleotid sorozat alakítja ki. A második és harmadik hajtű 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 hajtű külön-külön partnerrel és/vagy funkcióval rendelkezik, (ii) az első hajtű és a második-harmadik hajtű együttese külön feladatot lát el, (iii) mindhárom hajtű együttműködésben lát el egy egységes funkciót, (iv) a középső poli-A/U hajtű egy hőérzékelő struktúra (alacsony olvadásponttal rendelkezik) mely "mozgatja" a kétoldalán lévő első illetve harmadik hajtű 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 hajtűvel rendelkező másodlagos szerkezetet alakít ki. Ezt többféle szoftverrel is meghatároztuk; (A) ViennaRNAfold, (B) RNA composer, (C) McGenus.

4.2. K-mer analízis, hierarchikus klaszterelemzés

A K-mer analízist két céllal végeztük el; egyrészről, 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 régión kívül eső szakaszok hasonlóságát a cDNS és ncRNS adatok alapján, és (iii) amennyiben vannak hasonlóságok akkor azonosítsunk 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 Kmer 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 elemeztem.

4.2.1. 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 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 gének aktív eleme az RNS maga, nem pedig a DNS). Azért, hogy ezekből az esetleges eltérő kifejeződésekből adódó hamis következtetéseket elkerüljük, elvégeztük a K – mer analízist duplikátumok nélkül, összesen 142 cDNS/ncRNS szekvencián, a korábbiakhoz hasonlóan 3-6 K-mer értékekkel. Az eredmények megfeleltek a DNS szekvenciáknál tapasztaltaknak, azonban – mivel kevesebb faj kevesebb szekvenciájával dolgoztunk – 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 K - mer 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.

4.2.2. 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.



-0.49-0.32-0.160.01 0.17 0.34 0.5 0.67 0.83

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 (9. ábra).



^{-0.84-0.68-0.53-0.37-0.21-0.050.11 0.26 0.42 0.58 0.74 0.89 1.05 1.21 1.37 1.53 1.68 1.84}

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 legegyértelműbbek 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", ami alapján elképzelhető, hogy egy típusú funkciója van az összes HCR tartalmú homológnak.

4.2.3. A HCR régión kívüli szakaszok K-mer tartalmának vizsgálata

A cDNS/ncRNS szekvenciák nem-HCR régiónak 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 szakaszról is, aminek hossza 7-7000 nukleotidig változatos hosszúságú volt (11. ábra). A K – mer analízis eredményei alapján K – mer = 3-6 értékek esetén sem tudtunk jelentősen dúsuló K-mer mintázatot meghatározni a fajokon belül jellemző hasonlóságokon felül (12. ábra).



11. ábra: a cDNS és ncRNS szekvenciák HCR előtti és utáni hosszainak összehasonlítása A HCR-t tartalmazó cDNS és ncRNS szekvenciák HCR előtti és utáni szakaszainak hossza nagymértékű változatosságot mutatott, a HCR előtti, upstream szakaszok 76-2532 nt, a downstream szakaszok pedig 7-7000 nt hosszúságúak voltak.



12. 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).

4.2.4. A cDNS/ncRNS szekvenciák hieararchikus klaszterelemzése és a kialakult csoportok vizsgálata

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.

Az AtGUT15 és AtCR20 HCR-ek korábbi vizsgálataiból kiindulva, illetve a fajokban azonosított magas kópiaszámok (a DNS szekvenciák esetén 86 fajban azonosítottunk legalább 2 kópiát) alapján elképzelhető, hogy a HCR szekvenciák több, eltérő funkciókkal rendelkező csoportra bonthatók. Ahhoz, hogy ezt alátámasszuk, vagy éppen megcáfoljuk, a K-mer analízis eredményével, a Pearson – korrelációs mátrixszal dolgoztunk tovább (a K-mer analízist a teljes cDNS/ncRNS szekvenciákon, K-mer 3-6 értékekkel végeztük el). Kiválasztottuk az AtGUT15 (at2g18440) és AtCR20 (at4g36648) szekvenciák értékeit tartalmazó oszlopokat és ezeket ábrázoltuk egy olyan diagramon, aminek x tengelyét az AtGUT15, az y tengelyét pedig az AtCR20 jelentette. Az egyes értékek tehát azt mutatták meg, hogy az adott szekvencia K-mer ujjlenyomata mennyire hasonlít az AtGUT15 és mennyire az AtCR20 K-mer ujjlenyomatához (az 1-es érték jelenti a teljes egyezést). Ennek eredményeként azt tapasztaltuk, hogy az egyes szekvenciák értékeit (pár kivételtől eltekintve) középen, az AtGUT15 és AtCR20 szekvenciák értékeitől egyenlő

távolságra helyezkedtek el, nem alkottak tisztán elváló csoportokat K = 3-6 értékek esetén. Ebből arra következtettünk, hogy a K – mer analízis alapján nem válnak el tisztán különböző csoportok, tehát ha eltérő funkcióval is rendelkeznek ezt a funkciót nem a K – merek határozzák meg.



13. ábra: a szekvenciák K-mer ujjlenyomatának eloszlása az *AtGUT15 és AtCR20* függvényében A HCR-t tartalmazó szekvenciák K-mer ujjlenyomatait az *AtGUT15* (GUT15-ként jelölve) és *AtCR20* (CR20-ként jelölve) K-mer ujjlenyomataival összehasonlítva meghatároztuk, hogy ezek alapján nem válnak két hasonlósági csoportra (minél nagyobb az egyezés, az érték annál közelebbi az 1-hez).

Összesítve tehát, a HCR régiót tartalmazó cDNS és ncRNS szekvenciák HCR-en kívüli szakaszai mind hosszukban (11. ábra), mind K – mer ujjlenyomataikban (12. ábra) magas variabilitást

mutatnak. A K – mer analízis alapján nem alkotnak "GUT15 – szerű" és "CR20 – szerű" csoportokat (13. ábra), azonban valamiféle, a fajhatárokat átlépő hasonlóság felfedezhető bennük. Eredményeink azt sugallják, hogy a HCR régió szerkezete, nem pedig a szekvencia-mintázat (K-mer mintázat) lehet a funkcionális tulajdonság.

4.2.5. *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 hiearchikus 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 szerettük 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		er = 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.

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

Az lncRNS-ek definició szerint nem kódolnak fehérjét, azonban számos olyan lncRNS létezik mely mikro-peptid kódolásra képes (Choi és mtsai., 2019). Mivel a HCR régió önmagában egy konzervált szakasz, a HCR régiót átívelő rövid leolvasási keretekek (short Open Rearding 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 hajtűket összekötő szakaszok, a hajtűk csúcsi részének és a középső hajtű (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 hajtű nukleotid konzerváltságából egyenesen következik, de a továbbiakban nem fenntartott (14. ábra).

GUT15_ORF2: CR20_ORF2: BnaGUT15-chrA8_ORF2:	MSGALAWQVKKLILNKKKILW MTGALAWQVTK-ILNKKKKFR MTGALAWQVTL-NNNKNNCFV *:*******. **:::	VLERRSHGLLFFPGISS VVEGSSHGGFFLPGFSF LVRGVRTGDFFSPGSPF ::. * :* **	PLCVSSCLCSISLP-ALSHFIS LACVSCFSLFVHTTFDLDSFT- TLCLALSL *::	59 58 45
GUT15_ORF2: CR20_ORF2: BnaGUT15-chrA8_ORF2:	FLNAHIHSKTDHKQSL	75 58 45		

14. ábra: az AtGUT15, AtCR20 és BnaGUT15 (chrA8) gének HCR régióival átfedő sORF peptidek hasonlósági vizsgálata

Annak érdekében, hogy megállapítsuk, a HCR-ek funkciójukat egy mikro-peptiden keresztül fejtik-e ki, meghatároztuk a HCR-ek 10 aminosavnál hosszabb sORF-jeit, ezeket *in silico* transzláltuk és meghatároztuk konzerváltságukat a többi fajban.

Az Arabidopsis thaliana-ban azonosított sORF aminosav szekvenciája nem mutatott konzerváltságot, még a közeli rokon Brassica napus szekvenciájával sem (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 maga, amely valószínűleg másodlagos szerkezetén keresztül tudja ellátni funkcióját.

5. Következtetések

Munkánk során arra törekedtünk, hogy a korábbiaktól eltérő módszerekkel vizsgáljuk a GUT15/CR20 lncRNS családot és olyan információkhoz jussunk, amik a hagyományos módszerekkel nem elérhetők.

Ennek érdekében azonosítottuk az *AtGUT15* konzervált HCR régióját és ennek első hajtűje (első 37 nt hosszú szakasza)(3. ábra) alapján az általunk választott adatbázisban elérhető növényi DNS, cDNS és ncRNS adatokban meghatároztuk a HCR-el rendelkező szekvenciákat (4. ábra). Ezek nagy kópiaszáma (5. ábra) és konzerváltsága alátámasztja, hogy a GUT15/CR20 lncRNS-ek valamilyen funkcióval rendelkezhetnek. A továbbiakban meghatároztuk a HCR-ek konzerváltságát, illetve annak illeszkedését a fajok evolúciós változásaira. Ez alapján megállapítottuk, hogy a szekvenciák változásai nagy vonalakban illeszkednek a fajok evolúciós változásaira (6. ábra).

Az azonosított HCR-ek szerkezeti elemzése magas konzerváltságot mutatott az összes faj esetén, ami valószínűsíti, hogy a másodlagos, illetve harmadlagos szerkezet fontos eleme a HCR-ek működésének.

A HCR-ek további vizsgálatai során K – mer analízist és hierarchikus klaszterelemzést végeztünk mind a DNS, mind a cDNS/ncRNS szekvenciákon. Ennek eredménye egyfelől azt mutatja, hogy a HCR-ek K – mer ujjlenyomatai sok esetben fajokon átívelő hasonlóságokat mutatnak (8. ábra), másfelől a HCR-ektől upstream és downstream szakaszok olyan heterogenitással rendelkeznek (11., 12. ábra), ami valószínűsíti, hogy ezek a szakaszok a funkció kialakításában nem vesznek részt.

A K-mer mintázat alapján készült Pearson-korrellációs vizsgálatok nem különítettek el GUT-szerű és CR20-szerű funkcionális csoportokat (13. ábra). Továbbá, a *Triticum* fajok K – mer analízisének eredményei (8., 9., 10. ábra) a *Triticum aestivum* kialakulását figyelembevéve arra utalhatnak, hogy a HCR-ek az evolúció során uniformizálódhatnak. Ezek a megfigyelések azt valószínűsítik, hogy a GUT15/CR20 homologok egy funkcióval rendelkeznek, de ennek ellátásához azonban több kópia megléte szükséges.

Az sORF-ek azonosítása és konzerváltságuk elemzése alapján megállapítottuk, hogy a HCR-ek nem kódolnak olyan mikropeptideket, amiken keresztül funkciójuk kifejeződhet.

Következtetéseink tehát az alábbiak:

- 1. A GUT15/CR20 géncsaládban fellelhető HCR régió erőteljesen konzervált a törzsfejlődés során zárvatermőkben;
- 2. A GUT15/CR20 génekben megtalálható HCR régió egy konzervált másodlagos szerkezettel rendelkezik.
- 3. A HCR régión 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 kötést hoznak létre (töltésük révén), vagy a HCR stuktúra 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/transzkriptum közreműködésén alapul.
- 5. A k-mer mintátazának Pearson korrelációja és a Triticum sub-genomi GUT15/CR20 kópiák összehasonlító vizsgálata alapján egy típusú funkcionális csoportot feltételezünk a GUT15/CR20 lncRNS-eknek.
- 6. Mivel a GUT15/CR20 lncRNS-ek nem kódolnak konzervált sORF-et, feltételezésünk, hogy a funkcionális egység a HCR, mely aktivitásának alapja a másodlagos szerkezet.

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.

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 zárvatermők ősében jelent meg egy alkalommal és konzerválódott, és a zárvatermő 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 és Szaker Henrik Mihálynak a szakmai iránymutatásért és a közös munkáért.

8. Irodalomjegyzék

- Altschup, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. In *J. Mol. Biol* (Köt. 215).
- 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
- Batista, P. J., & Chang, H. Y. (2013). Long noncoding RNAs: Cellular address codes in development and disease. In *Cell* (Köt. 152, Szám 6, o. 1298–1307). Elsevier B.V. https://doi.org/10.1016/j.cell.2013.02.012
- Bonasio, R., & Shiekhattar, R. (2014). Regulation of transcription by long noncoding RNAs. Annual Review of Genetics, 48, 433–455. https://doi.org/10.1146/annurev-genet-120213-092323
- 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
- Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., ... Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9), 1775–1789. https://doi.org/10.1101/gr.132159.111
- 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., Yatusevich, 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
- Flynn, R. A., & Chang, H. Y. (2014). Long noncoding RNAs in cell-fate programming and reprogramming. In *Cell Stem Cell* (Köt. 14, Szám 6, o. 752–761). Cell Press. https://doi.org/10.1016/j.stem.2014.05.014
- Gabory, A., Jammes, H., & Dandolo, L. (2010). The H19 locus: Role of an imprinted non-coding RNA in growth and development. In *BioEssays* (Köt. 32, Szám 6, o. 473–480). https://doi.org/10.1002/bies.200900170
- Geisler, S., & Coller, J. (2013). RNA in unexpected places: Long non-coding RNA functions in diverse cellular contexts. In *Nature Reviews Molecular Cell Biology* (Köt. 14, Szám 11, o. 699–712). https://doi.org/10.1038/nrm3679
- 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
- Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports*, 11(7), 1110–1122. https://doi.org/10.1016/j.celrep.2015.04.023
- 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
- Jian Ye, Scott McGinnis, & Thomas L. Madden. (2006). BLAST: imporvements for better sequence analysis. *Nucleic Acids Research*, 34(Web Server).
- Jin, J., Liu, J., Wang, H., Wong, L., & Chua, N. H. (2013). PLncDB: Plant long non-coding RNA database. In *Bioinformatics* (Köt. 29, Szám 8, o. 1068–1071). https://doi.org/10.1093/bioinformatics/btt107
- 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
- Kelley, D., & Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. http://genomebiology.com/2012/13/11/R107
- 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
- Lauret Durent, Corinne Chureau, Sylvie Samain, Jean Weissenbach, & Philip Avner. (2006). The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene. *Science*, *312*(5780), 1650–1653. https://doi.org/10.1126/science.1126863
- Li, L., Eichten, S. R., Shimizu, R., Petsch, K., Yeh, C. T., Wu, W., Chettoor, A. M., Givan, S. A., Cole, R. A., Fowler, J. E., Evans, M. M. S., Scanlon, M. J., Yu, J., Schnable, P. S., Timmermans, M. C. P., Springer, N. M., & Muehlbauer, G. J. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biology*, 15(2). https://doi.org/10.1186/gb-2014-15-2-r40
- Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., & Chua, N. H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in arabidopsisC W. *Plant Cell*, 24(11), 4333–4345. <u>https://doi.org/10.1105/tpc.112.102855</u>
- Lu, C., MacIntosh, G., Green, P. (2003). Two Non-coding RNAs, AtGUT15 and AtCR20, modulate the ABA response at high temperature in Arabidopsis., <u>Publication Detail</u> (arabidopsis.org)

- Marques, A. C., & Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: Modest conservation and incompleteness. *Genome Biology*, 10(11). https://doi.org/10.1186/gb-2009-10-11-r124
- 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
- Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. In *Human molecular genetics: Köt.* 15 Spec No 1. https://doi.org/10.1093/hmg/ddl046
- Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., & Mattick, J. S. (2008). *Specific* expression of long noncoding RNAs in the mouse brain. www.pnas.org/cgi/content/full/
- 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
- Pattanayak, D., Solanke, A. U., & Kumar, P. A. (2013). Plant RNA Interference Pathways: Diversity in Function, Similarity in Action. In *Plant Molecular Biology Reporter* (Köt. 31, Szám 3, o. 493–506). https://doi.org/10.1007/s11105-012-0520-9
- Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan, L., Sandelin, A., Rinn, J. L., Regev, A., & Schier, A. F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research*, 22(3), 577– 591. https://doi.org/10.1101/gr.133009.111
- 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
- Pin Wang, Yiquan Xue, Yanmei Han, Li Lin, Cong Wu, Sheng Xu, Zhengping Jiang, Junfang Xu, Qiuyan Liu, & Xuetao Cao. (2014). The STAT3-Binding Long Noncoding RNA Inc-DC Controls Human Dendritic Cell Differentiation. *Science*, 344(6181), 307–310. https://doi.org/10.1126/science.1250897

- 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 tRNAlike 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
- Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. In *Cell* (Köt. 136, Szám 4, o. 629–641). Elsevier B.V. https://doi.org/10.1016/j.cell.2009.02.006
- Quinn, J. J., & Chang, H. Y. (2016a). Unique features of long non-coding RNA biogenesis and function. In *Nature Reviews Genetics* (Köt. 17, Szám 1, o. 47–62). Nature Publishing Group. https://doi.org/10.1038/nrg.2015.10
- Quinn, J. J., & Chang, H. Y. (2016b). Unique features of long non-coding RNA biogenesis and function. In *Nature Reviews Genetics* (Köt. 17, Szám 1, o. 47–62). Nature Publishing Group. https://doi.org/10.1038/nrg.2015.10
- 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
- Rinn, J. L., & Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual Review* of *Biochemistry*, 81, 145–166. https://doi.org/10.1146/annurev-biochem-051410-092902
- Romano, G., Veneziano, D., Acunzo, M., & Croce, C. M. (2017). Small non-coding RNA and cancer. In *Carcinogenesis* (Köt. 38, Szám 5, o. 485–491). Oxford University Press. https://doi.org/10.1093/carcin/bgx026
- 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
- Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., & Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*, 41(17). https://doi.org/10.1093/nar/gkt646
- Sunkar, R., Li, Y. F., & Jagadeeswaran, G. (2012). Functions of microRNAs in plant stress responses. In *Trends in Plant Science* (Köt. 17, Szám 4, o. 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.
- Ulitsky, I., & Bartel, D. P. (2013). XLincRNAs: Genomics, evolution, and mechanisms. In *Cell* (Köt. 154, Szám 1, o. 26). Elsevier B.V. https://doi.org/10.1016/j.cell.2013.06.020
- Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., & Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147(7), 1537–1550. https://doi.org/10.1016/j.cell.2011.11.055
- Wang, H., Chung, P. J., Liu, J., Jang, I. C., Kean, M. J., Xu, J., & Chua, N. H. (2014). Genomewide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Research*, 24(3), 444–453. https://doi.org/10.1101/gr.165555.113
- 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
- Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., & Deng, X. W. (2014). Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28), 10359–10364. https://doi.org/10.1073/pnas.1409457111
- Wierzbicki, A. T., Haag, J. R., & Pikaard, C. S. (2008). Noncoding Transcription by RNA Polymerase Pol IVb/Pol V Mediates Transcriptional Silencing of Overlapping and Adjacent Genes. *Cell*, 135(4), 635–648. https://doi.org/10.1016/j.cell.2008.09.035
- Wright, E. S. (é. n.). TITLE 1 RNAconTest: Comparing tools for non-coding RNA multiple sequence alignment based on 2 structural consistency 3 Running title: RNAconTest: benchmarking comparative RNA programs 4. http://DECIPHER.codes/Downloads.html
- 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 heatinducible 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

KONZULTÁCIÓS NYILATKOZAT

A Gál Nóra (hallgató Neptun azonosítója: JKUW77) konzulenseként nyilatkozom arról, hogy a diplomadolgozatot áttekintettem, a hallgatót az irodalmi források korrekt kezelésének követelményeiről, jogi és etikai szabályairól tájékoztattam.

A diplomadolgozatot a záróvizsgán történő védésre javaslom / nem javaslom.

A dolgozat állam- vagy szolgálati titkot tartalmaz: igen <u>nem</u>

Kelt: Gödöllő, 2023 május 3.

Cyrch Tiber

Belső konzulens

NYILATKOZAT

a diplomadolgozat nyilvános hozzáféréséről és eredetiségéről

A hallgató neve:	Gál Nóra
A Hallgató Neptun kódja:	JKUW77
A dolgozat címe:	A hőindukált GUT15/CR20 lncRNS család vizsgálata bioinformatikai módszerekkel
A megjelenés éve:	2023
A konzulens tanszék neve:	Növénybiotechnológia tanszék

Kijelentem, hogy az általam benyújtott diplomadolgozat egyéni, eredeti jellegű, saját szellemi alkotásom. Azon részeket, melyeket más szerzők munkájából vettem át, egyértelműen megjelöltem, s az irodalomjegyzékben szerepeltettem.

Ha a fenti nyilatkozattal valótlant állítottam, tudomásul veszem, hogy a Záróvizsga-bizottság a záróvizsgából kizár és a záróvizsgát csak új dolgozat készítése után tehetek.

A leadott dolgozat, mely PDF dokumentum, szerkesztését nem, megtekintését és nyomtatását engedélyezem.

Tudomásul veszem, hogy az általam készített dolgozatra, mint szellemi alkotás felhasználására, hasznosítására a Magyar Agrár- és Élettudományi Egyetem mindenkori szellemitulajdon-kezelési szabályzatában megfogalmazottak érvényesek.

Tudomásul veszem, hogy dolgozatom elektronikus változata feltöltésre kerül a Magyar Agrár- és Élettudományi Egyetem könyvtári repozitori rendszerébe.

Kelt: 2023 május 3

Haligató aláírása