Doktori (PhD) értekezés tézisei

Humán transzkripciós faktorok összehasonlító elemzése ChIP-seq adatokkal és transzkripciós faktorok komplex topológiai elrendeződésének vizsgálata a DNS-en

Czipa Erik

Témavezető: Dr. Barta Endre

DEBRECENI EGYETEM
Molekuláris Sejt- és Immunbiológia Doktori Iskola

Debrecen, 2019

Humán transzkripciós faktorok összehasonlító elemzése ChIP-seq adatokkal és transzkripciós faktorok komplex topológiai elrendeződésének vizsgálata a DNS-en

Értekezés a doktori (PhD) fokozat megszerzése érdekében
a klinikai orvostudományok tudományágban

Írta: **Czipa Erik**
okleveles Molekuláris Biológus

Készült a Debreceni Egyetem Molekuláris Sejt- és Immunbiológia Doktori Iskola keretében

Témavezető: Dr. Barta Endre, PhD

A doktori szigorlati bizottság:

elnök: Prof. Dr. Bíró Sándor, az MTA doktora
tagok: Dr. Vámosi György, PhD
Dr. Gáspári Zoltán, PhD

A doktori szigorlat időpontja: Debreceni Egyetem ÁOK, Biokémiai és Molekuláris Biológiai Intézet
2018. március 22. 10 óra

Az értekezés bírálói:

Dr. Gáspári Zoltán, PhD
Dr. Penyige András, PhD

A bírálóbizottság:

elnök: Prof. Dr. Fésüs László, az MTA rendes tagja
tagok: Dr. Gáspári Zoltán, PhD
Dr. Penyige András, PhD
Dr. Benkő Szilvia, PhD
Dr. Sebestyén Endre, PhD

Az értekezés védésének időpontja: Debreceni Egyetem ÁOK, Belgyógyászati Intézet „A” épület tanterme2020. január15. 13 óra

1. BEVEZETÉS

A genomi DNS tartalmazza a szervezet felépítéséhez szükséges legfontosabb információkat. A szervezet közel minden sejtjében megtalálható ennek a másolata, ami humán esetében több trillió kópiát jelent. A szervezetet felépítő sejtek morfológiai sokszínűségének hátterében a különböző genomi régiók aktvitása illetve inaktivitása áll. A gének expressziója befolyásolja a sejtek proteomját, mely funkcionális sajátosságokkal ruházza fel a különböző típusú sejteket. A gének expresszója egy több szinten szabályozott folyamat, melyet meghatároz többek között a kromatin (DNS-fehérje komplex) szerkezete, illetve transzkripció folyamata.

Transzkripció reguláció és általános transzkripciós faktorok

A génexpresszió központi eseménye a transzkripció, melyben a genetikai információról RNS másolatot készítenek az RNS-polimeráz enzimek. A szintézis "engedélyezése" több lépcsős folyamat. Az RNS prekurzok létrejötte előtt, a szabályozásban visszafelé haladva, az első folyamat a pre-iniciációs komplex (PIC) összeszerelődése, melyet az általános transzkripciós faktorok (general transcription factors (GTF)) alkotnak. Ezek a fehérjék a transzkripciós start helyen (TSS), illetve a promóter régióban (amik a gén 5' végén helyezkednek el) található specifikus DNS szekvenciákat, a központi promóter elemeket (DPE, TATA-box, TCT, BRE és MTE) ismerik fel. A specifikus DNS szekvenciákat felismerve kötődnek a PIC megfelelő elemei, ezt követően a komplex többi tagja is a TSS-hez rekrutálódik. A komplex összeszerelődése után az RNS-polimeráz II. (POLII) elkezdi a pre-RNS szintézist. Ennek zavartalan működéséhez a kromatin megfelelő konformációja szükséges. Ez a transzkripciós faktorok közvetítésével jön létre.

Transzkripciós faktorok és kofaktorok

GTF-eken kívül még más, több mint 3230 féle különböző transzkripciós faktort is megkülönböztetünk (TF). Ezek a DNS-hez kapcsolódó GTF-ekhez hasonlóan rendelkeznek DNS kötő doménnel, mellyel specifikus szekvenciát kötnek meg, de nem képezik a PIC részét. Ezek a fehérjék kollaborációban más faktorokkal (további TF-ek vagy kofaktorok (CF)) aktiválják vagy represszálják a transzkripciót. Ezt leggyakrabban a kromatin kondenzáltsági állapotának módosításával teszik lehetővé: hiszton fehérjék poszttranszlációs módosítása és nukleoszómák újra rendezése (kromatin újra modellezés), topológiailag asszociált domének

(TAD) létrehozása és rendezése, illetve al-TAD (sub-TAD) domének módosítása (kromatin hurkolódás).

Topológiailag asszociált domének

A DNS-t ha egyenessé hajtogatnánk akkor egy két méter hosszú szállat kapnánk. Ez a szál egy legfeljebb 5 μm átmérőjű sejtmagba csomagolódik be. A csomagolódás első szintje a nukleoszóma, azaz a DNS szál (146 bázispáronkénti) körbetekerődése a hiszton fehérjéken. Ezek körülbelül egy 11 nm gyöngyfüzér struktúrát hoz létre, mely tovább csomagolódik a nukleuszban. Az interfázisban lévő sejtmagban az örökítő anyag kevésbé kondenzált aktív (eukromatin) és kondenzált inaktív (heterokromatin) részre oszlik.
Ezeknek a létrehozásában jelentős szerepet játszanak a 3C (kromoszóma konformáció rögzítés) technikákkal ismertté vált kromoszóma domének, az úgynevezett topológiailag asszociált domének (TAD-ok). Ezek száz kilobázistól néhány millió bázispár hosszú DNS szakaszokat is magukba foglalhatnak, melyek a kromoszómákat különböző funkcionális egységekre osztják. Szerepük még kevéssé ismert, viszont különböző vizsgálatokkal már bizonyították a régiókon belüli autonóm gén regulációt, melyben a TAD-ban található gének hasonló expressziós profillal rendelkeznek. Ezek az evolúciósan konzervált régiók több sejtosztódást követően is konstansnak bizonyulnak. A TAD-ok további részekre, az úgynevezett al-TAD (sub-TAD) doménekre tagolódnak. Ezek olyan DNS hurkok, melyekben disztális DNS régiók (anchor - horgony régiók) kerülnek közel egymáshoz. Az interfázisban lévő sejtben a hurkok horgany régióit a CTCF és kohezin gyűrű komplex tartja köti össze.

CTCF és kohezin

A CTCF egy 11 cink-ujj doménnel rendelkező fehérje. A cink-ujjak felelősek a specifikus "CCCTC" DNS szekvencia kötéséért, melyről a fehérje a nevét kapta. A fehérje képes interakcióba lépni a kohezin gyűrű tagjaival (RAD21, SMC1/2 és STAG1), melyek egy ~50 nm szélességű hajtű struktúrát hoznak létre. Ez képes körbefogni a DNS-t és a távoli DNS régiókat közel tartani egymáshoz.

Transzkripciós faktorok vizsgálata nagy áteresztő képességű szekvenálási technikákkal

A funkcionális genomika és molekuláris biológia mérföldkövének számít a nagy áteresztőképességű szekvenálási technikák megjelenése (High-Throughput Sequencing (HTS)). A HTS segítségével lehetővé vált a relatív könnyű és gyors párhuzamos szekvenálás.

A HTS kombinációja más molekuláris biológiai eljárásokkal lehetővé teszi az epigenetikai, genomikai és transzkriptomikai kutatások globális szintre emelését. Feltérképezhető sejtpopulációk transzkriptomja (RNA-seq), nukleoszóma mentes régiói (DNase-seq, ATAC-seq), DNS interakciós hálózata (HiC; ChIA-PET) és DNS-fehérje interakciója (ChIP-seq).
Dolgozatomban elsősorban a ChIP-seq technikára fókuszálok, mely egy hatékony technika a különböző DNS-fehérje interakciós helyek azonosítására. A kromatin immunprecipitáció (ChIP) során a fehérjéket kereszt kötjük egymással és a DNS-el. Random fragmentálást követően a fehérje-DNS komplexet a vizsgálni kívánt fehérje ellenes anitesttel kezeljük, majd "kihalásszuk" ezeket a keresztkötött komplexeket az oldatból. A szekvenálást és számítógépes feldolgozást követően globális képet kapunk egy bizonyos transzkripciós faktor vagy kofaktor genomi lokalizációiról. A formaldehides keresztkötésnek köszönhetően nem csak közvetlenül a DNS-hez kapcsolódó transzkripciós faktorokat vizsgálhatjuk, hanem azokat a fehérjéket is, amelyek egy komplexen keresztül, közvetve csatlakoztak a DNS szálhoz (kofaktorok).
Transzkripciós faktor és kofaktor ChIP-seq esetében a feldolgozást követőn a fehérjék által elfoglalt genomi helyeket úgynevezett "csúcs" ("peak") régióként detektáljuk, ahol a legmagasabb jelintenzitású pont, az úgynevezett csúcspont ("summit") képviseli a fehérje valódi lokalizációját.

2. CÉLKITŰZÉS

Munkacsoportunk korábban kimutatta, hogy a CTCF és kohezin gyűrű tagjainak (RAD21, SMC1/2, STAG1) vizsgálatánál, hogy a fehérjék ChIP-seq csúcspontjai között szál specifikus elcsúszás látható, ami követi a CTCF motívum irányát. Mivel a komplex tagjai közül csak a CTCF rendelkezik DNS kötő doménnel, ezért csúcspontoknak azonos pozícióra kellene esniük. A fehérjék csúcspontjai jól meghatározott sorrendet mutatnak egymáshoz viszonyítva, ami fehérje topológiai sajátosságokra utal. Feltételezzük, hogy a csúcspont pozíciók egymás mellé rendeződése a kohezin fehérjék topológiai elhelyezkedésére utal.
Számot egér és humán ChIP-seq adatot töltöttünk le hogy választ kapjunk a következő kérdésekre:

- Minden mintánél megfigyelhető e az előzőekben megfigyelt elcsúszás a CTCF és kohezin csúcspontok között?
- Az elcsúszás jellemezhető e szál specificitással?
- Az elcsúszás milyen összefüggésben áll a CTCF motívum orientációjával?
- Megfigyelhető e valamilyen sorrend a komplexet alkotó fehérjék ChIP-seq jelei között?

- Figyelhető-e meg korreláció a fehérjék mért pozícióija és a CTCF-kohezin komplex röntgen krisztallográfiai szerkezete között?
- A fehérjék pozíciója milyen összefüggésben áll a CTCF mediálta kromatin hurkolódással?

Ezt követően kibővítettük vizsgálatainkat más transzkripciós faktorokra is. Nyilvános adatbázisokból a lehető legtöbb elérhető humán ChIP-seq adatot próbáltuk összegyűjteni. Ezeket elemezve adatbázist készítettünk az eredményekből, mellyel a következőeket akartuk elérni:

- Egy automatikus ChIP-seq elemző program sort készíteni, mely az adatokat egységesen dolgozza fel és képes a fehérjék topológiai elhelyezkedésének vizsgálatára.
- Egy olyan adatbázist létrehozni, mely nemcsak a transzkripciós faktorok genom szintű kötőhelyeit tartalmazza, de a rajta elhelyezkedő fehérjék pontos pozícióját is.
- Felfedezni olyan új fehérje-fehérje interakciókat a transzkripciós faktorok között, melyek eddig ismeretlenek voltak.
- AZ adatokat nyilvánosan élhetővé kívántuk tenni egy web felület segítségével.

3. ANYAGOK ÉS MÓDSZEREK

Adatgyűjtés

A National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) adatbázisából több mint 4068 humán ChIP-seq adatot gyűjtöttünk össze. Ezeknek az adatoknak a meta-adatait egy általunk fejlesztett perl script elemezte: elnevezte, majd automatikusan szűrte a mintákat. Elsődleges tervünk egy olyan adatbázis létrehozása volt, ahol zavartalanul össze lehet hasonlítani az adatokat. Ezért kiszűrtük a géncsendesítésen/génkiütésen átesett sejtvonalakat.

Elsődleges analízis

A ChIP-seq adatok feldolgozásához egy, a munkacsoportunk által fejlesztett scriptet használtunk, mely tartalmazza többek között a szekvenált DNS fragmentek ("read"-ek) térképezését a referencia genomhoz (hg19/GRCh37 referencia genom BWA algoritmussal), a csúcsok prediktálását és de novo motívum analízist (HOMER program csomag).

Peakek szűrése és elválasztása

A csúcspont alapú topológiai analízisekhez jól meghatározható csúcspontokra volt szükségünk. Ehhez a fals pozitív csúcsok minimálisra csökkentésére és a több csúcsponttal rendelkező régiók elválasztására volt szükség. Utóbbihoz a PeakAnalyser csomag PeakSplitter programját használtuk, mely külön régiókra tagolta a vállal és több csúcsponttal rendelkező csúcsokat. Ezt követően ezeket a régiókat szűrtük egy általunk fejlesztett perl programmal. Ez a csúcsok lefutását és alakját figyelembe véve próbálja csökkenteni a fals pozitívan prediktált ChIP-seq kötőhelyeket/genomi lokalizációkat.

Motívum optimalizáció és transzkripciós faktor kötőhely azonosítás

A motívumok genomra való térképezéséhez a JASPAR CORE adatbázis motívum mátrixait használtuk. Ez 579 nem redundáns motívumot jelent, melyek irodalmi adatokból származnak. A motívumokat optimalizáltuk a maximális kötőhely azonosítás érdekében. Az összegyűjtött és feldolgozott ChIP-seq adatokat használtuk fel a kötőhelyek genomi pozíciójának azonosításához. Minden JASPAR CORE motívumhoz megpróbáltunk megfelelő ChIP-seq kísérletet párosítani (transzkripciós faktorra specifikus ChIP-seq-hez az irodalmi adatokban azonosított kötőhelyüket pl.: CTCF motívum CTCF csúcs; RXR:RAR kötőhely- RXR és RAR ChIP-seq csúcs) és ezek csúcs régióit felhasználva optimalizáltuk a motívumokat, majd kerestünk a genomban kötőhelyeket (az optimalizált motívumokat használva). Így az azonosított kötőhelyek a ChIP-seq kísérleteknek köszönhetően validáltak. 338 JASPAR CORE motívumhoz sikerült humán ChIP-seq kísérletet párosítani. 3 programot használtunk az optimalizált motívumok térképezéséhez: HOMER annotatePeaks.pl, FIMO és MAST. A konszenzus motívum szettbe azokat a motívumokat válogattuk, amelyeket legalább 2 program azonosított. A redundáns kötőhely azonosítás elkerüléséhez a csúcspontok alapján szelektáltuk. Azok a pozíciók kerültek a végleges adatsorba, amelyek centrumához legközelebb volt a ChIP-seq-el azonosított fehérje csúcspont (a motívumot kötő fehérjéé).

Csúcspont távolság és motívum centrum távolság kalkuláció

Az adattáblák készítéséhez megvizsgáltuk az azonosított kőtőhelyek és a velük átfedésben lévő ChIP-seq kísérleti eredményeket. Ezzel képet kaphattunk arról, hogy a különböző kötőhelyekhez a különböző sejtvonalakban, milyen fehérjék rekrutálódhatnak. A kötőhelyeket és azok centrumát referencia pontként használtuk a fehérje távolság kalkulációban, mivel fix pontot képviselnek a genomon. A BEDTools szoftver csomag programjaival megkerestük a

kötőhelyek közelében találhat csúcspontokat, minden ChIP-seq kísérletből. A csúcspont-motívum centrum távolságokat szál specifikus módon számoltuk ki és bázispárban mértük.

4. EREDMÉNYEK

CTCF és kohezin fehérjék topológiai sorrendje

A CTCF és kohezin alegységek ChIP-seq csúcsai és csúcspontjai között elcsúszás figyelhető meg. A CTCF-kohezin komplexben csak a CTCF rendelkezik DNS kötő doménnel és a többi fehérje rajta keresztül közvetve kapcsolódik a DNS szálhoz. Ennél fogva a csúcspontok megközelítőleg ugyanazon a genomi pozíción való elhelyezkedését várnánk. A CTCF kötőhelyek behatóbb vizsgálata rámutat az elcsúszás szál specifikusságára. A CTCF kötőhelyét referenciaként használtuk a csúcspont- motívum távolság mérésénél, melyben a motívum középső bázisa volt a viszonyítási pont. Mivel a CTCF egy nem palindrom "CCCTC" szekvenciát képes kötni, ezért a meghatározott konszenzus kötőhely szett motívumaihoz szál specificitást tudtunk rendelni. 237 humán és 183 egér kohezin és CTCF ChIP-seq adatot dolgoztunk fel. Az azonosított CTCF kötőhelyeken megvizsgáltuk a különböző kísérletek fragmentum lefedettségét, ami a ChIP-seq jelek intenzitásának feleltethető meg. Az eredményeket grafikusan ábrázoltuk és azt kaptuk, hogy az átlagos lefedettség maximum pozíciója (a ChIP-seq jelintenzitás legmagasabb pontja (a CTCF kötőhely középpontjához viszonyítva)) a fehérjék között szegregációt mutat. A szegregáció minden sejtvonalban ugyanazt a meghatározott sorrendet követi a fehérjék között ami a motívum mentén 5'-3' irányba haladva a következő: CTCF > SMC1/3 > RAD21 > STAG1. Ez nem követi az eddig publikált interakciós sorrendjét a proteineknek: CTCF > STAG1 > RAD21 > SM1/3. A pozíció analízist jobban kibontva összesítettük a csúcspont- motívum centrum távolságokat. A kísérleteket szeparáltuk sejtvonal szerint és ennek megfelelően összehasonlítottuk a CTCF adatokat a kohezin adatokkal. A csúcspont-motívum centrum távolságokat eloszlás hisztogramokkal és boxplotok segítségével ábrázoltuk. A különböző fehérjék átlag pozícióját a viszonyítási ponthoz képest bázispár pontossággal meg tudtuk határozni. A CTCF fehérje a kötőhely centrumhoz képes átlagosan ~ -4 bp, az SMC1/3 ~1 bp, a RAD21 6 bp és a STAG1 ~7 bp pozíciónál volt megtalálható. A fehérjék pozíció különbségének szignifikaciáját párosított t-próbával ellenőriztük. A P érték $P<10^{-15}$ volt Wilcoxon és Friedman teszt alapján.
Fehérje struktúra archívumokban (RCSB PDB) megtalálhatóak a CTCF és a kohezin alegységek szerkezetének részletei. Ezeket az adatokat, a fehérjék interakciós sorrendjét és a saját adatainkat felhasználva készítettük egy hipotetikus modellt a CTCF mediálta kromatin

hurkolódásról. A modell készítése során feltételezzük, hogy a CTCF és STAG1-RAD21 pozíció közötti ~11 bázispáros távolságot a molekulák szerkezeti sajátosságai, illetve a formaldehides keresztkötés okozza. Egy DNS fordulatnak a hosszúsága megközelítőleg 11 bp. A RAD21 és STAG a CTCF-hez kapcsolódnak, azonban komplex struktúrájának köszönhetően fizikai közelségbe kerülnek a legközelebbi DNS régióval (ami a CTCF jellel szomszédos DNS kis-árok közelében található) és kromatin immunprecipitáció során a fehérjék és a DNS között kereszt-kötés jön létre. Az SMC1/3 fehérjék szokatlan helyzetét a kettős-gyűrű ("double-embrace") modellel tudjuk magyarázni. Ebben a modellben a kromatin hurkot létrehozó CTCF-fehérjékhez mind a két oldalon kapcsolódik 1-1 kohezin gyűrű. Ezek az ellentétes oldali DNS szálat ölelik körbe és hozzák fizikai közelségbe a disztális DNS régiókat. Az SMC1/3 fehérjék hinge doménnel rendelkeznek. Ezek felelősek többek között az SMC1 és SMC3 közötti heterodimerizációért, illetve nem specifikus DNS kötő szerepüket is leírták már. Ha az általunk kapott eredményeket vizsgáljuk, a B-DNS modellre térképezett fehérje pozíciókon szembetűnő, hogy míg a CTCF-RAD21-STAG1 a DNS azonos oldalára rendeződnek, addig az SMC1/3 jelek ennek az ellentétes oldalán találhatóak. Ezért feltételezzük, hogy a detektált jel a hurkolódásban részt vevő ellentétes oldali CTCF-kohein komplex SMC1/3 fehérjéihez tartozik. A kohezin hajtű CTCF-től távolabb eső oldalán található hinge domén a hurok ellentétes oldalán lévő DNS szálhoz nem specifikus módon kapcsolódik, illetve formaldehid kezelés hatására kereszt-kötődik ezzel a DNS régióval. Az általunk készített hipotetikus modell a kettős gyűrű elméletet támogatja.

Számítógépes szimuláció

A csúcseltolódási értékek validálásához számítógépes szimulációt végeztünk, hogy megvizsgáljuk mekkora elemszám szükséges ahhoz, hogy az általunk kapott eltolódási értékek reprodukálhatóak legyenek. A HeLa sejtvonalból származó adatokból (n=21994) több körben random kiemeltünk értékeket. A kiemelt értékek elemszámát minden körben megnöveltük. Az eredményeket logaritmikus skálán ábrázoltuk, melyből kiderül, hogy legalább 100 kötőhely vizsgálatával reprodukálhatóvá vált az általunk kapott eredmény +/- 1 bp pontossággal. A csúcspont alapú pozíció meghatározásnak tehát a detektáláshoz szükséges esetszám alsó határa 100.

Kísérleti validálás

A kísérleti validálásnál az irodalomból már ismert heterodimerizációs vagy interakciós partnereket használtunk. Pozitív kontrol estén olyan fehérjéket vizsgáltunk, melyek

rendelkeznek kompozit elemmel, vagy közös kötőhelyük ismert. A csúcspont-kompozit elem centrum közötti távolságok eloszlás görbéi a két fehérje esetében jól elkülöníthető maximumot mutatnak egymás mellett, ami követi a kompozit elemben lévő kötőhelyek sorrendjét. Erre a FOXA1 transzkripciós faktort és az androgén receptort használtuk. VCap és LNCap sejtvonalakból származó ChIP-seq kísérleteket elemeztünk és az előzőekben már ismertetett módon ábrázoltuk. A hisztogramokon a fehérjék pozíciója követte a kompozit elemekben található kötőhelyek sorrendjét, ezért a validálás sikeres volt.

Negatív kontrolként olyan fehérje komplexet vizsgáltunk, ahol 2 vagy több fehérje közül csak 1 rendelkezik DNS kötő doménnel. Mivel az egyik fehérje a másikon keresztül, közvetve csatlakozik a DNS-hez, azt vártuk, hogy a két fehérje pozíciója ugyanarra a genomi pozícióra térképeződik. Erre a célra a P300-RXR (3T3-L1 sejtvonalból származó ChIP-seq adatok) és a NANOG-POU5F1 (egér embrionális őssejtből származó ChIP-seq adatok) példáját használtuk fel. Az RXR rendelkezik ismert DNS kötő doménnel, míg a P300 kofaktorként közvetlenül vagy közvetve az RXR-hoz csatlakozik. NANOG-POU5F1 esetében csak a POU5F1 rendelkezik DNS kötő doménnel. A referencia pontok ebben az esetben DNS kötő doménnel rendelkező fehérjék kötőhelyei voltak. Mind a két esetben az interakciós párok átlagos pozíciója ugyanazon a genomi ponton volt megtalálható.

CTCF mediálta kromatin hurkolódás

A CTCF-kohezin vizsgálati eredmények tükrében megvizsgáltuk, hogy a fehérjék topológiai sorrendje milyen összefüggésben állhat a CTCF mediálta DNS hurkokkal. Ehhez CTCF ChIA-PET adatokat töltöttünk le nyilvános adatbázisokból (ENCODE, NCBI SRA) és megvizsgáltuk az interakcióban résztvevő genomi régiókat. A ChIA-PET által azonosított interakciós régiók alacsony felbontása megnehezíti a hurkot kialakító CTCF kötőhelyek azonosítását. Az általunk készített konszenzus CTCF kötőhelyeket felhasználva (melyek ChIP-seq kísérletekkel voltak validálva) megpróbáltuk a lehető legtöbb ChIA-PET interakció fehérje kötőhelyét azonosítani. Ezt követően páronként megvizsgáltuk az azonosított motívumok orientációját (melyik szálon található a CCCTC motívum) az interakció mind a két oldalán. A lehetséges variációk közül (konvergens +-; divergens -+, azonos szálon lévő motívumok -- vagy ++) a konvergens orientáció bizonyult dominánsnak, azaz a CTCF motívumok egyike a pozitív a másik a negatív szálon volt megtalálható, a két motívum mindegyike a hurok belseje felé irányult. Az eredmények konzisztensek voltak minden vizsgált sejtvonalban (K562, mouse limbbud, MCF7).

Hiszton módosítások vizsgálata a CTCF mediálta kromatin hurkokban

A TAD-on belül található kromatin hurkok kohezin asszociáltak és olyan funkcionális egységeket formálnak, mint például az enhaszer-promoter interakciók. Ezek gyakran asszociáltak CTCF fehérjével is, melyek pontos funkcióját nehéz meghatározni: egyszerre működhetnek gén aktivátorként illetve inzulátorként is.

Ezek vizsgálatához töltöttünk le és elemeztünk hiszton módosításokat célzó ChIP-seq kísérleteket és vizsgáltuk a jelintenzitásukat a CTCF mediálta kromatin hurkok horgony régióin. Az előzőekben definiált MCF7 interakciós szettet használtuk (azonosított CTCF kötőhelyekkel) 7 nyilvános adatbázisban elérhető MCF7 ChIP-seq adattal (H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K27ac, H3K27me, H3K36me). Megvizsgáltuk az interakciókat létrehozó CTCF kötőhelyeken a különböző hiszton ChIP-seq adatok átlagos jelintenzitását. Azt tapasztaltuk, hogy a H3K4 metilációs adatok különösen magas fragment lefedettséget mutattak ezeken a régiókon.

Ezt követően szeparáltuk azokat az interakciókat, melyek egyik oldala mutatott átfedést H3K4 metilációs ChIP-seq csúcsokkal. Ez a korábban azonosított 7234 interakcióból 4188-at érintett. Az összes interakciós ponton vizsgáltuk a H3K4 metilációs mintázatokat (a CTCF kötőhely 100 bázispáros kereteiben) és az eredményt k-Means klaszterezéssel csoportosítottuk. Az eredményt heatmap segítségével ábrázoltuk. A legszembetűnőbb jelenség az interakciók oldalai között megfigyelhető aszimmetria. Az interakciók túlnyomó többsége az egyik oldalán magas jelintenzitást mutatott valamely H3K4 metiláció jelet tekintve, míg a másik oldal alacsonyt. A klaszterezés eredményeként 3 csoportot tudtunk elkülöníteni:

- 1. klaszter: erős H3K4me2 és H3K4me3 jel, alacsony H3K4me1
- 2. klaszter: magas H3K4me1, közepes H3K4me2 és alacsony H3K4me3 jel
- 3. klaszer: magas H3K4me2, közepes H3K4me1 és alacsony H3K4me3

Az annotációs elemzés megmutatta, hogy az 1. klaszterbe tartozó interakciós régiók promoter specifikus hurkolódásban vesznek részt, míg a 3. és a 2. klaszterbe tartozók enhaszer-enhanszer illetve enhaszer-intron hidakat képeznek.

ChIPSummitDB

A CTCF-kohezin vizsgálata során kifejlesztett csúcspont alapú topológia predikciót más transzripciós faktorok elemzésénél is alkalmaztuk. Ezért egy átfogó transzkripciós faktor kötőhely és fehérje topológiai adatbázist készítettünk. Az ehhez szükséges programokat és szkripteket egy 1500 kísérletet tartalmazó minta adatszetten fejlesztettük ki. A kezdeti fázis

után ezt az adatszettet kibővítettük. Nyilvános adatbázisból (NCBI SRA) 4052 humán ChIP-seq adatot gyűjtöttünk össze, melyeket az általunk kifejlesztett pipeline-al elemeztünk. 3782 ChIP-seq kísérletet elemeztünk sikeresen. A kötőhelyek azonosításához a JASPAR CORE adatbázis motívumait használtuk. 338 motívumhoz tudtunk ChIP-seq kísérletet párosítani. A párosítatlan motívumok olyan irodalmi forrásból származnak, ahol a forrásként használt adatok fasta file-jait nem tették elérhetővé, vagy nem rendelkeznek emberből származó ChIP-seq adattal. A letöltött és sikeresen elemzett 3782 ChIP-seq adat közül 2496 tartozott olyan transzkripciós faktorhoz, melynek a JASPAR CORE adatbázisában szerepel ismert motívuma. A maradék ChIP-seq adatok jelentős része kofaktorokhoz tartozik.

A motívumok 1/7-ét érinti a palindromitás problémája. Ez megnehezíti a szál specifikus fehérje pozíció-motívum centrum távolság kalkulálását, mivel a motívum mind a két szálon ugyanabban a pozícióban található. Ezt a motívum 5' oldalához kapcsolt szegélyező bázisok hozzáadásával próbáltuk megoldani. Ennek lényege, hogy további bázispárokat kapcsolunk a motívum egyik oldalához (NNN szekvenciák, amik bármilyen bázist jelölhetnek), aminek következtében a motívum centruma eltolódik. A két szálon lévő motívumok közepe a genom nem ugyanazon pontjára esik (referencia pontok). A két motívum közül az kerül kiválasztásra, amelyiknek a centrumához közelebb található a saját fehérjéjéhez tartozó ChIP-seq csúcspont. Ezzel a módszerrel a bilaterális motívumok arányát sikerült 30 % alá csökkentenünk.

Az azonosított transzkripciós faktor kötőhelyeket referencia pozícióként használtuk a csúcspont távolságok kalkulálására. Megvizsgáltunk minden azonosított transzkripciós faktor kötőhelyet és megnéztük milyen fehérjék, milyen sejtvonalból mutatnak ChIP-seq jelet a közelében s ezeknek mi a preferált pozíciója. Ezekből átlagot számoltunk és a ChIPSummitDB webfelületen vizsgálható az átlagtávolság a motívum középpontjától, de lehetőség van a pontos eloszlás görbe vizsgálatára is.

Az eredményekből MySQL segítségével adatbázist készítettünk, mely elérhető a http://summit.med.unideb.hu/summitdb/ weboldalon keresztül.

Adatbázis és webes felület

A ChIPSummitDB-n keresztül a következő adatok érhetőek el:

- Optimalizál transzkripciós faktor motívumok: a JASPAR motívumok, melyek az NCBI SRA adatbázisából letöltött ChIP-seq kísérletekkel lettek optimalizálva.
- Transzkripciós faktor kötőhelyek genomi lokalizációja: Az optimalizált motívumok a nekik megfelelő ChIP-seq csúcsok figyelembevételével lettek visszatérképezve a referencia genomra (hg19/GRCh37).

- Fehérje hálózat adatok: minden transzkripciós faktor kötőhely és a vele átfedő ChIP-seq kísérletek listája.
- Topológiai adatok: A különböző fehérjék preferált pozíciója a motívumok centrumához viszonyítva, csúcspont alapú elemzésből kalkulálva.
- ChIP-seq kísérlet meta-adatok: minden kísérlet részletes leírása és tulajdonságai, megfelelő linkekkel az adat forrásához.
- Szabályozó régiókban található SNP-k: a különböző transzkripciós faktor kötőhelyekkel átfedő SNP-k, melyek megzavarhatják a fehérje kötőhely felismerését.

Ezek az adatok nem csak letölthetők, hanem a weboldalon a különböző megjelenítési módokban vizualizálhatóak is: MotifView, ExperimentView, PairShiftView, GenomeView és dbSNPView.

MotifView:

A MotifView-ban egy összefoglaló ábrát kaphatunk a kiválasztott traszkripciós faktor kötőhely típusról (például: CTCF, RXR, RXR:RAR, ATF3 stb.) és azokról a ChIP-seq kísérletekről, amelyeknek a csúcspontja megtalálható volt a motívum centrum +/- 50 bázisparos keretén belül. Ez a scatterplot típusú megjelenítési mód megadja a ChIP-seq kísérletek átlagos motívum centrum-csúcspozíció távolságát és ezek elemszámát illetve szórását. Minden pont egy ChIP-seq kísérletet képvisel és a pontok a cél-fehérjéjük alapján vannak színezve. Így könnyen vizsgálható, hogy a különböző fehérjék milyen pozíciót preferálnak.
Ha a távolság értékek szórását nézzük, látható, hogy a különböző típusú fehérjék szegregációt mutatnak, a proteinek különböző "rétegekbe" rendeződnek. Ha közelebbről megvizsgáljuk a jelenséget láthatjuk, hogy a szórás egyenes arányban áll a fehérje DNS-től való távolságával. Azok a transzkripciós faktorok, melyek a válaszadó elemüknél közvetlenül kapcsolódnak a DNS-hez alacsonyabb pozíció szórást mutatnak. Azoknál a fehérjéknél, melyek ehhez a proteinhez kapcsolódnak, vagy egy komplexen keresztül csatlakoznak a kötőhelyhez magasabb szórás figyelhető meg.

PairShift View

Míg a MotifView egy átfogó képet mutat a különböző kísérletek pozíció preferenciájáról, a PairShiftView ennél részletesebb információt ad 1-3 kísérlet lokalizációs eloszlásáról. Akárcsak a MotifView esetében az X tengelyen a kiválasztott motívumtól való távolság. Egy hisztogramon nyomon lehet követni, hogy a kiválasztott kísérletek csúcspontjai milyen gyakran fordulnak elő a különböző pozíciókban a motívum +/- 50 bázisparos keretén belül.

ExperimentView

Az elemzett kísérletek legfontosabb információit foglalja össze. Ez a nézet tartalmazza a kísérlethez tartozó: eredeti forrást, az összes read számot, az azonosított csúcsok számát, kísérleti azonosítókat, az adott kísérlet mely motívumokkal mutat átfedést, transzkripciós faktorok esetében a hozzá tartozó motívumokat és a de novo motívumkeresés eredményeit. Utóbbi megmutatja az összes feldúsult motívumot a kísérletben azonosított csúcs régiók alatt.

Genome view

Ebben a nézetben megközelítőleg az összes adatunk vizsgálható egy genom böngészőben. Akár az analízis során, itt is a referencia genom a hg19. Egy kétdimenziós nézetben vizsgálhatjuk a kísérletekben azonosított csúcs és csúcspont régiókat. Megjeleníthetjük az azonosított transzkripciós faktor kötőhelyeket.

dbSNPview

A dbSNP egy publikus domén a humán nukleotid-polimorfizmusok archiválására. Ezt az adatbázist integráltuk, hogy vizsgálni lehessen a válaszadó elemekkel átfedő SNP-ket. A transzkripciós faktor kötőhelyeket és SNP pozíciókat itt is egy kétdimenziós genom böngészőhöz hasonló nézetben jelenítjük meg. A kötőhelyek PWM-ként vannak megjelenítve, így meg tudjuk vizsgálni, hogy az SNP a motívum melyik bázisára esett, az mennyire esszenciális a fehérje-DNS interakcióban. Ezzel következtetni lehet arra, hogy a különböző rSNP-k milyen mértékben zavarhatják meg a fehérje kapcsolódását.

CTCF kötőhelyek kapcsolata a génexpresszió szabályozással

A CTCF kötőhelyekről általánosan elmondható, hogy a magas számban fednek át különféle transzkripciós faktor ChIP-seq csúcsokkal. Különösen szembetűnő, hogy nem csak kohezin fehérjék ChIP-seq jelei csoportosulnak a CTCF motívum 3' végére, hanem az összes többi faktoré is. Megvizsgáltuk a CTCF kötőhelyekkel leggyakrabban átfedő transzkripciós faktorokat és a kohezin fehérjék mellett a YY1 és ZNF143 bizonyult a leggyakoribbnak. Ez általános érvényű a különböző sejttípusok között. Sejtvonalanként vizsgáltuk az összetartozó CTCF, RAD21, YY1, ZNF143 és SMC1/3 ChIP-seq jeleket CTCF kötőhelyeken és hierarchikus (Manhatten távolság) klaszterezés után heatmap-en ábrázoltuk az eredményeket. Ebből kiderült, hogy egyéb faktor csúcsok csak azokon a CTCF kötőhelyeken találhatóak, amelyek YY1 és ZNF143 csúcsokkal is átfedtek. Ez az eredmény a négy vizsgált sejtvonalban (GM12878, H1hESC, HeLA és K562) konzisztensnek bizonyult. Tovább szeparáltuk a CTCF

kötőhelyeket YY1-nal átfedő és nem átfedő régiókra és megvizsgáltuk a faktorok előfordulási gyakoriságát. Az eredmény megerősítette a heatmap-nél megfigyelteket, YY1 hiányában a transzkripciós faktorok szinte teljesen eltűntek, míg YY1 jelenlétében jól látható feldúsulás figyelhető meg.

A ZNF143 és YY1 csúcspont pozícióit tekintve azonos genomi pontra térképeződnek az SMC1/3 fehérjékkel. A megfigyelés érdekes, hiszen azt sugallja, hogy a YY1 és a ZNF143 az SMC-k hinge doménjéhez képesek kapcsolódni. Ha a pozíciók szórását vizsgáljuk, akkor azt láthatjuk, hogy a ZNF143 alacsonyabb szórással rendelkezik a YY1-nál. Ez az előzőekben leírtak alapján egy közvetlenebb kapcsolódást jelent a kohezin gyűrűhöz. Ez nem meglepő, mivel a ZNF143-at írták le, mint a kohezin gyűrű egyik alkotóelemét. A YY1 alacsonyabb szórásából azt feltételezzük, hogy a ZNF143-on keresztül kapcsolódik a kohezin gyűrűhöz. A többi transzkripciós faktor alacsony ChIP-seq jel intenzitása, illetve a pozíciók magas szórása ennél még távolibb kapcsolatra utalhat, miszerint komplexen keresztül kötődhetnek a régióhoz. Megvizsgáltuk a CTCF és a YY1-ZNF143 csúcsok közötti átfedéseket. Mind a három fehérjének figyelemreméltóan sok helyen található ChIP-seq jele a genomon. GM12878 sejtvonalban ábrázoltuk az átfedéseket Venn diagramon. A metszetekre tagolt csúcs régió szettek is jelentős elemszámot mutatnak. A különböző metszetek elemeit megvizsgáltuk, hogy milyen arányban fednek át CTCF kötőhelyekkel. A CTCF csúcsok körülbelül 55 %-a rendelkezik azonosított kötőhellyel. Ez az arány ZNF143 illetve YY1 esetében jelentősen alacsonyabb. A CTCF-el átfedést mutató ZNF143 és YY1 kötőhelyek motívum aránya is jelentős. A fehérjék számos helyen lefedik a genomot. A ChIP-seq eredmények alapján nem csak a saját kötőhelyen mutatnak lefedettséget, hanem olyan régiókon is amiket látszólag közvetlenül nem tudnak kötni, a detektált jel követett interakció eredménye lehet más fehérjéken keresztül. CTCF esetében az indirekt kölcsönhatást a ZNF143, illetve a YY1-nal való kapcsolat hozhatja létre.

Az előző eredményekből feltételezzük, hogy az enhaszer-promoter interakciók létrehozásában a kohezin gyűrű passzívan vehet részt és a transzkripciós faktor által kötött promoter és kohezin gyűrű között a ZNF143 és a YY1 képezhet hidat.

GATA1 és TAL1 kötőhelyek vizsgálata

Korábbi publikációkban már leírták a GATA1 és TAL1 transzkripciós fakorok közötti kooperációt eritroid differenciáció során. A két fehérje juxtapozíciója olyan gyakori, hogy a JASPAR CORE adatbázisban is szerepel a kompozit elemük. Az LMO2 komplex mediátor szerepet tölt be a két fehérje közötti kapcsolat létrehozásában.

Ha külön vizsgáljuk a GATA1 fehérjék csúcs távolság eloszlását a GATA1 kötőhelyeken, akkor azt tapasztaljuk, hogy a gyakoriság eloszlás görbe maximuma a névadó GATAA szekvenca kezdő (G) bázispárjához képest 7 bázispárral eltolva található meg a 3' irányban. A konszenzus GATA1 motívum centrumtól ez 9-10 bázispárra található. Ez a megfigyelés érvényes az összes GATA (GATA1, GATA2, GATA3, GATA4 és GATA6) fehérjére bármely sejttípusból.

TAL1 esetében a csúcspozíciók maximuma nem határozható meg pontosan. A TAL1 és TCF3 (E2A) hélix-hurok-hélix doménen keresztül kötik a DNS-t és heterodmerizációjuk jól ismert. Kompozit elemük is nyilvántartott a JASPAR CORE adatbázisban. Mind a két fehérje egy-egy CAG szekvenciához képes kapcsolódni, ami a heterodimerizációt követően a genomon egy CAGCTG régiót jelent. Ennek a palindromitása megnehezíti a fehérjék pozíciójának pontos meghatározását, a referencia pont és szál információ kijelölésének nehézségei miatt.

Mivel a GATA1:TAL1 kompozit elem a GATAA szekvencia jelenléte miatt nem számít palindromnak, ezért a TAL1 motívumnak is egyszerűbben tudjuk meghatározni a szál specificitását. Ha a kompozit kötőhelyet vizsgáljuk, a TAL1 csúcspozíciók eloszlás görbéjének a maximuma a TAL1 kötőhely "CA" szekvenciájánál található.

A scatterploton jól megfigyelhető a GATA1 és TAL1 szegregációja a kompozit elem mentén. Az eloszlás hisztogramokon megfigyelhető maximumok is jól elkülöníthetően egymástól. A fehérjék pozíciójának a sorrendje követi a kompozit elemben a GATA1 és a TAL1 motívumok pozícióját.

A GATA1 fehérje elcsúszását a kötött motívumhoz képest a fehérje szerkezeti sajátosságaiban kerestük. A GATA1 proteinnek röntgen krisztallográfiai módszerekkel a 15 %-át sikerült eddig feltérképezni (63 aminosav a 413-ból). Az ismeretlen fehérje régiók jelentős hányada az N terminális oldalon található (~55 %). Emellett található a már ismert szerkezetű DNS kötő cink-ujj domén. A peptidlánc C-terminális részén található a protein fennmaradó ismeretlen ~29 %-a. B-DNS modellen vizsgálva az ismert fehérjeszerkezetet azt láthatjuk, hogy az ismeretlen szerkezetű N-terminális oldal abban az orientációban helyezkedik el, ahol a GATA1 csúcspozíciók maximumát prediktáltuk. Feltételezzük, hogy a fehérje fennmaradó része fizikailag közel kerülhet a DNS azon részéhez mely egy fordulattal a GATA1 motívum 3' irányában található. ChIP-seq preapárció során ezen DNS szakasz között és a fehérje között keresztkötés alakulhat ki formaldehid kezelés után. Ennek köszönhető a GATA1 csúcspont eltolódása a kötött motívumhoz képest.

Szabályozó régiókban található SNP-k vizsgálata a dbSNPView segítségével

A szekvencia variációk a szabályozó régiókban a transzkripciós faktorok kötődését befolyásolhatják, ami génexpressziós változásokhoz is vezethet. Ezeknek az úgynevezett regulatórikus SNP-knek (rSNP) az azonosítása jelentős kihívást jelent. Nem csak a variációk azonosítása jelent nehézséget, hanem a szabályozó régióké is. Mivel az adatbázisunk szabályozó régiók gyűjteménye, ezért kézenfekvő volt egy rSNP kereső funkció beépítése. Integráltuk a dbSNP adatbázist, amely több mint 893 millió humán SNP gyűjteménye. Ezeket egy genom böngészőhöz hasonló felületen lehet vizsgálni a velük átfedést mutató transzkripciós faktor kötőhelyekkel. A kötőhelyek PWM-ként való megjelenítése megkönnyíti a következtetést az SNP zavarására a faktor kötődése során. A kötési eseményben részt vevő "erős" bázisokon található variációk nagyobb eséllyel zavarják meg a transzkripciós faktor szekvencia felismerését.

Az úgynevezett dbSNPView-t használva specifikus kötőhelyeket vagy SNP-ket kiválasztva vizsgálhatjuk a genomot rSNP-k után kutatva. Hogy validáljuk a dbSNPView-t, olyan variációkat kerestünk melynek publikálták a génexpresszitót befolyásoló hatását. Így találtuk meg az rs2742624 SNP-t, mely az UPK3A gén enhanszer régiójában található. Az UPK3A fehérje a kiválasztó rendszer felépítésében játszik szerepet. A gén csökkent expressziója renális diszpláziához vezet. Egy nemrég megjelent tanulmányban vizsgálták az rs2742624 UPK3A gén expressziójára kifejtett hatását. Mivel az SNP egy GATA kötőhelyben található, csökkentette a GATA fehérjék affinitását. Ennek következtében a gén expressziója jelentősen represszálódott. Az általunk azonosított GATA kötőhelyek szintén átfedést mutatnak az SNP-vel. Adataink alpján nem csak a GATA2, hanem a GATA3, GATA4 és GATA5 fehérjék kötődését is befolyásolhatja a pontmutáció.

5. DISZKUSSZIÓ

A molekuláris biológia és a szekvenálás az utóbbi évtizedekben ugrásszerűen fejlődött. Ez a szekvenálási adatok felhalmozódásához vezetett. Ezen adatok tárolását, megosztását és hivatkozhatóvá tételét tűzték ki célul olyan adatbázisok, mint az NCBI SRA, ENCODE vagy DDBJ. A nyers adatok számos feldolgozási lépésen esnek át míg végül vizsgálhatni lehet őket. A processzálás kihívást jelentő feladat, melyben a módszerek kiválasztása jelentősen befolyásolhatja az eredményeket.

A szekvenálási adatokban habár óriási potenciál rejlik, az általuk kínált információk csupán töredéke kerül részletes kivizsgálásra. A legtöbb labor többnyire egy problémára fókuszáltan

végzi el vizsgálatait, és csak kevés munkacsoport rendelkezik olyan nagy számítógépes erőforrással, hogy vállalja ezen adatok nagy mennyiségű, egységes feldolgozását.

Vizsgálataink során a ChIP-seq adatok feldolgozására fókuszáltunk. Korábbi vizsgálati eredményeinkből kiderült, hogy a ChIP-seq nem csak a fehérjék megközelítőleges pozíciójának meghatározása használható a genomon, hanem fehérje-topológiai információk is kinyerhetőek belőlük. Kifejlesztettünk egy úgynevezett csúcspont pozíció vizsgálati módszert, melyben referencia pontokat (fix genomi pozíciót) használva meghatározóható a fehérjék helyzete egy állandó genomi pozícióhoz, majd egymáshoz viszonyítva. A technikát a CTCF-kohezin komplex vizsgálatával teszteltük. Ezt követően a módszert nagyobb léptékben használva humán ChIP-seq adatokat töltöttünk le, ügyelve arra, hogy a vizsgált fehérjék és sejtvonalak minél szélesebb választékát megvizsgáljuk. Az adatokból adatbázist készítettünk, amelyet egy nyilvános web-es felületen keresztül elérhetővé tettünk.

A CTCF és kohezin (RAD21, STAG1, SMC1/3) komplex DNS hurkokat hoznak létre interfázisban lévő sejtekben. A komplex tagjai közül a CTCF feladata, hogy közvetlenül a DNS-hez kapcsolódjon. 11 cink ujj doménjével egy specifikus CCCTC szekvenciát ismer fel és kapcsolódik hozzá. A kohezin gyűrű és a CTCF közötti interakcióért a CTCF C terminális része és a STAG1 felelős. A CTCF mintegy "kihorgonyozza" a kohezin gyűrűt a megfelelő genomi pozícióknál. A komplex többi tagja struktúrális szerepet tölt be a hurok létrehozásában. Főként a SMC1/3 fehérjék tekercselt tekercs szerkezete egy ~50 nm átmérőjű hajtűt hoz létre, mely képes a DNS átfogására és két disztális régió fizikai közelségbe hozására egy hurok formálásával. A komplex elemeinek ChIP-seq eredményeit vizsgálva elcsúszásra lehetünk figyelmesek a CTCF és a kohezin gyűrű tagjai között. Mivel a komplex tagjai közül csak a CTCF képes a DNS közvetlen kötésére, ezért a ChIP-seq csúcspontok elhelyezkedését minden fehérjénél ugyanazon a genomi pozíción várnánk. Az elcsúszást azonban közelebbről megvizsgálva azt tapasztaltuk, hogy az szál specifikus.

Hogy mérhetővé tegyük az elcsúszást és annak szabályszerűségeit referencia pontot használtunk. Erre a feladatra kézenfekvő volt a CTCF kötőhely szál specifikus használata (referencia szekvencia CCCTC). A kötőhely centrumához viszonyítva a különböző fehérjék átlagos csúcspont elhelyezkedését egy meghatározott sorrendiség jellemzi. A motívum közepétől negatív irányba találhatóak a CTCF ChIP-seq jelek átlagai és maximumai, majd szál specifikusan a pozitív irányba (5' > 3') haladva a komplex többi tagjai a következő sorrendben: SMC1/3, RAD21 és STAG1. A faktorok csúcspont elhelyezkedését átlag pozíciókká konvertáltuk és B-DNS modellen ábrázoltuk. Ezen jól látható az SMC1/3 szokatlan elhelyezkedése, ami a B-DNS modellen a CTCF-RAD21-STAG1 fehérjékkel ellentétes oldalon

található. A már meglévő fehérje szerkezeti adatokat felhasználva hipotetikus modellt készítettünk a CTCF mediálta kromatin hurkolódásról. A modellünk az újnevezett "double-embrace", kettős gyűrű elméletet támogatja, mellyel magyarázható az SMC1/3 pozíciója is. Az SMC fehérjék úgynevezett hinge doménnel rendelkeznek, mely a hajtű ellentétes oldalán található a komplex többi tagjához képest. Ez a domén aspecikus interakció révén kötődhet a DNS hurok ellentétes oldalához.

ChIA-PET adatok segítségével vizsgáltuk meg, hogy a fehérjék topológiai sorrendje milyen összefüggésben állhat a CTCF mediálta DNS hurkokkal. AZ ENCODE adatbázisában található ChIA-PET adatokat használtuk fel (MCF7, K562 és HeLa). Az analízis során a konvergens motívum irányultság bizonyult dominánsnak (~70 % az MCF7 sejtvonalban), melyben az egyik motívum a pozitív szálon, a másik a negatív szálon található, így a CCCTC motívum a hurok belseje felé orientálódik. Ennek megfelelően a kohezin gyűrű is a hurok belsejében található.

Az MCF7 sejtvonalból származó CTCF ChIA-PET kísérletek replikáiból készítettünk egy konszenzus kromatin hurok szettet. Ezt összehasonlítottuk az adatbázisban rendelkezésünkre álló MCF7 hiszton ChIP-seq eredményekkel. Az eredmények alapján a konszenzus szett 60%-a volt összefüggésbe hozható valamely aktív promotert/enhanszert jelölő hiszton módosítással (H3K4me, H3K4me2, H3K4me3). Ez összefüggésbe hozható a későbbi YY1-ZNF143-al kapcsolatos vizsgálati eredményeinkkel.

A CTCF kötőhelyek jelentős átfedést mutatnak az egyéb transzkripciós faktorokat/ kofaktorokat célzó ChIP-seq kísérletekben azonosított csúcs régiókkal. Ezek közül is a legjelentősebb átfedést minden vizsgált sejtvonal esetében a YY1 és a ZNF143 mutatta a kohezin alegységek mellett. A jelenséget hierarchikus klaszterezéssel (Manhattan távolság) tovább vizsgálva kiderült, hogy a CTCF kötőhelyeken csak YY1 és ZNF143 jelenlétében figyelhető meg egyéb faktor ChIP-seq jel (például RUNX3, COREST, MAX stb.). A CTCF kötőhelyeket két populációra bontva vizsgáltuk: YY1-al átfedő és nem átfedő CTCF kötőhelyek. Az eredmény megerősítette a korábbi megfigyeléseket. YY1 jelenlétében jelentős feldúsulás figyelhető meg az átfedő faktorok számában, míg annak hiányában ezeknek az átfedéseknek a gyakorisága közel 90 %-al csökken. Ezek az eredmények szintén konzekvensek az összes vizsgált sejtvonalban (GM12878, MCF7, HeLa, K562).

A ZNF143 és YY1 átlagos csúcspont helyét vizsgálva azt láthatjuk, hogy az SMC1/3 fehérjékkel azonos pozíciót preferálnak. Globális analízis során megfigyeltünk, hogy a különböző kötőhelyeken a faktorok jól látható "rétegekbe" rendeződnek az pozíciók szórását vizsgálva. A scatterplot ábrázolás során a fehérjék nem csak a preferált pozíciójuk alapján szegregálódnak, hanem a pozíció szórásuk alapján is. Jobban megvizsgálva a jelenséget azt

tapasztaltuk, hogy a szórás egyenes arányban állhat a faktor-DNS proximitással. Egy faktornak annál alacsonyabb a pozíció szórása, minél közelebb helyezkedhet el a DNS-hez. Így a legelső "réteget" a motívumot közvetlenül kötődő gazda transzkripciós faktorok alkotják. Ezt követik az ehhez kapcsolódó kofaktorok vagy indirekt kötő transzkripciós faktorok, majd a rekrutált komplex többi tagja. Hasonlóan különbség mutatkozik a CTCF és YY1-ZNF143 között. A CTCF relatív alacsony pozíció szórást (~17) mutat a többi faktorhoz képest. A következő rétegben találhatóak a kohezin komplex tagjai a ZNF143-al (SD:~20). Ezek felett található a YY1 az egyéb transzkripciós faktorok/kofaktorok értékeivel. Mivel a ZNF143-at már leírták, mint a kohezin gyűrű tagját, ezért közvetlen kapcsolódást feltételezünk a ZNF143 és a kohezin alegységek között. A fehérjék pozíciójából és szórásából arra következtetünk, hogy a ZNF143 az SMC1/3 hinge doménjéhez képes kötődni, és hozzá pedig a YY1 fehérje. Utóbbi hidat képezhet a hurok belsejében található transzkripciós faktorokkal, melyek a szabályozó régióhoz kapcsolnak, ezzel segítve a promóter-enhanszer interakciót.

A csúcspont alapú topológiai elemzést kibővítettük egyéb faktorokra. 3782 humán ChIP-seq adatot dolgoztunk fel és az eredményekből elkészítettük a ChIPSummitDB adatbázist és weboldalt. A ChIPSummitDB egyéb feldolgozott ChIP-seq adatokat tartalmazó adatbázisokkal (mint például a GTRD, ChIPBase, ChIP-Atlas) több szempontból átfed:

- Nagy mennyiségű adat kollekciót kínál
- Egységesen feldolgozott ChIP-seq eredményeket tartalmaz
- ChIP-seq csúcs predikciós eredményeket tartalmaz
- Letölthető tartalom
- Kombinálható fájl formátumok
- ChIP-seq eredmények de novo motívum keresési eredményeit tatalmazza

Emellett számos csak a mi adatbázisunkra jellemző funkcióval rendelkezik. A legtöbb adatbázis a lehetséges kötőhelyek azonosítását végzi el a ChIP-seq csúcspozíció prediktálással. Kevés munkacsoport vállalja a valódi transzkripciós faktor kötőhelyek azonosítását, a motívumok genomi pozíciójának meghatározását. A ChIPSummitDB nemcsak rendelkezik ilyen információval, de a fehérje pozíció elemzések ezeken a régiókon alapszanak, mivel referencia pontként használtuk őket. Emellett az eredményeink nem csak letölthetően, hanem eloszlás diagrammok, scatterplotok és genom böngésző segítségével a weboldalunkon grafikusan megjeleníthetőek. A fehérje pozíciókat a GATA1:TAL1 kompozit elem vizsgálatával validáltuk, melyben a fehérjék pozíciója követte a meglévő a kompozit elem motívumainak orientáltságát és egybehangzó eredményt mutattak a röntgen krisztalográfiai eredményekkel.

Az integrált SNP keresővel azonosítottunk olyan regulatórikus SNP-ket melyek megzavarják a génexpressziót és klinikai jelentőséggel bírnak.

6. ÖSSZEFOGLALÁS

Kimutattuk, hogy a ChIP-seq technika megfelelő feldolgozással fehérje topológiai információk kinyerésére is használható. Egy ChIP-seq csúcspont alapú módszert dolgoztuk ki a fehérjék helyzetének meghatározásához egy fix genomi pozícióhoz viszonyítva. A technikát a CTCF-kohezin komplex vizsgálatával fejlesztettük. Ennek eredményéként meghatároztuk a CTCF és a kohezin gyűrűk helyzetét és készítettünk egy hipotetikus modellt a CTCF mediálta kromatin hurkokról. Meghatároztuk, hogy a hurok formálásban résztvevő CTCF motívumok konvergens orientációja a kohezin gyűrűt is interior elhelyezkedésre kényszeríti. Vizsgáltuk a CTCF mediálta hurkok transzkripciós faktorokkal és kofaktorokkal történő interakcióját és közvetett szerepét a génexpresszió szabályozásban. Eredményeink a YY1 és a ZNF143 lehetséges mediátor szerepére világítottak rá a kohezin gyűrű és a faktorok által kötött szabályozó régiók között.

Az elemzéseinket kibővítettük a lehető legtöbb transzkripciós faktor vizsgálatával. 3782 humán ChIP-seq adatot dolgoztunk fel egységesen és adatbázist készítettünk az eredményekből. Ezeket nyilvánosan elérhetővé tettük egy webfelületen, melyet ChIPSummitDB-nek neveztünk el (http://summit.med.unideb.hu/summitdb). Ez az eredményeinket nem csak letölthetővé, hanem ábrázolhatóvá is teszi. Ezen web eszközök segítségével vizsgáltuk a GATA1:TAL1 kompozit elemen a kötő fehérjék pozícióját. Ennek eredményei egybevágtak a már publikált fehérjestruktúra és egyéb kísérleti eredményekkel.

Finanszírozás

Ennek a projektnek a pénzügyi támogatását a GINOP-2.3.2-15-2016-00044, a 2017-1.3.1-vke-2017-00026 és a FIKP_20428-3_2018_FELITSTRAT biztosította.

DEBRECENI EGYETEM
EGYETEMI ÉS NEMZETI KÖNYVTÁR
H-4002 Debrecen, Egyetem tér 1, Pf.: 400
Tel.: 52/410-443, e-mail: publikaciok@lib.unideb.hu

Nyilvántartási szám: DEENK/376/2019.PL
Tárgy: PhD Publikációs Lista

Jelölt: Czipa Erik
Neptun kód: SYFW19
Doktori Iskola: Molekuláris Sejt- és Immunbiológia Doktori Iskola

A PhD értekezés alapjául szolgáló közlemények

1. **Czipa, E.**, Schiller, M., Nagy, T., Kontra, L., Steiner, L., Koller, J., Pálné, S. O., Barta, E.: ChIPSummitDB: a ChIP-seq based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them.
   Database. [Epub ahead of print], 2019.
   DOI: http://dx.doi.org/10.1093/database/baz141
   IF: 3.683 (2018)

2. Nagy, G., **Czipa, E.**, Steiner, L., Nagy, T., Pongor, S., Nagy, L., Barta, E.: Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA.
   BMC Genomics. 17 (637), 1-9, 2016.
   DOI: http://dx.doi.org/10.1186/s12864-016-2940-7
   IF: 3.729

DEBRECENI EGYETEM
EGYETEMI ÉS NEMZETI KÖNYVTÁR
H-4002 Debrecen, Egyetem tér 1, Pf.: 400
Tel.: 52/410-443, e-mail: publikaciok@lib.unideb.hu

További közlemények

3. Simándi, Z., **Czipa, E.**, Horváth, A., Kőszeghy, Á., Bordás, C., Póliska, S., Juhász, I., Imre, L., Szabó, G., Dezső, B., Barta, E., Sauer, S., Károlyi, K., Kovács, I., Hutóczki, G., Bognár, L., Klekner, Á., Szűcs, P., Bálint, B. L., Nagy, L.: PRMT1 and PRMT8 regulate retinoic acid-dependent neuronal differentiation with implications to neuropathology.
Stem Cells. 33 (3), 726-741, 2015.
DOI: http://dx.doi.org/10.1002/stem.1894
IF: 5.902

A közlő folyóiratok összesített impakt faktora: 13,314
A közlő folyóiratok összesített impakt faktora (az értekezés alapjául szolgáló közleményekre): 7,412

A DEENK a Jelölt által az iDEa Tudóstérbe feltöltött adatok bibliográfiai és tudománymetriai ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2019.11.20.

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *