# ChIP-seq-piikkien koostaminen ydinestimoinnilla

Pauli Tuoresmäki

Pro gradu -tutkielma

ITÄ-SUOMEN YLIOPISTO

Tietojenkäsittelytieteen laitos

Tietojenkäsittelytiede

Marraskuu 2015

ITÄ-SUOMEN YLIOPISTO, Luonnontieteiden ja metsätieteiden tiedekunta, Kuopio
Tietojenkäsittelytieteen laitos
Tietojenkäsittelytiede

Opiskelija, Pauli Tuoresmäki: ChIP-seq-piikkien koostaminen ydinestimoinnilla
Pro gradu -tutkielma, 59 s., 1 liite (1 s.)
Pro gradu -tutkielman ohjaajat: FT Wilhelmiina Hämäläinen ja FT, Dos. Sami Heikkinen
Marraskuu 2015

Tiivistelmä: Biolääketieteen tutkimuksen painopiste on pikkuhiljaa siirtynyt laboratoriosta datan analysointiin. Yksi syy tähän ovat ChIP-seq:n kaltaiset menetelmät jotka tuottavat paljon dataa kohtalaisilla kustannuksilla ja työmäärillä. ChIP-seq:llä voidaan tutkia proteiinin sitoutumista DNA:han. Yksi ChIP-seq-koe tuottaa miljoonia merkkijonoiksi koodattuja sekvenssilukemia. ChIP-seq-kokeista saadut raakadatat on yleensä julkaisujen yhteydessä annettava saataville julkisiin tietokantoihin, mikä on lisännyt käytettävissä olevan datan määrää. Datan määrästä johtuen useaa eri ChIP-seq-koetta samanaikaisesti hyödyntäville menetelmille olisi kysyntää. Sitoutumispaikan selvittämistä varten raakadata pitää kuitenkin jalostaa ChIP-seq-piikeiksi. Pelkkiä yhden ChIP-seq-kokeen piikkejä voidaan käyttää tutkimuksissa hyväksi, mutta usean eri kokeen samanaikainen hyödyntäminen voi tuottaa enemmän tietoa. Koetulosten yhdistäminen on kuitenkin osoittautunut haasteelliseksi. Menetelmät ovat siksi keskittyneet lähinnä parittaisiin vertailuihin ja varsinaisia yhdistämis-/koostamismenetelmiä ei ole juuri kehitetty. Tämän tutkimuksen tarkoituksena oli selvittää miten ChIP-seq-piikkejä voitaisiin koostaa ja voisiko koostamisessa käyttää apuna ydinestimointia. Selvityksen perusteella kehitettiin uusi ydinestimointiin perustuva ChIP-seq-piikkien koostamismenetelmä, ConsensusSummit. Menetelmässä käytetty ydinestimointi on tiheysfunktion estimointimenetelmä, jonka avulla voidaan käyttää piikkien tiheyttä koostamisen perusteena. Usean eri ChIP-seq-kokeen piikkejä käytettäessä tiheimpien alueiden piikit yhdistyvät koostepiikiksi ja harvassa olevat piikit pysyvät erillään. ConsensusSummit-menetelmää testattiin julkisella datalla, jolle oli saatavilla vertailukelpoisia tuloksia. Vertailun lisäksi testauksessa pyrittiin määrittämään menetelmälle sopivia parametreja. Lisäksi tutkittiin menetelmän kykyä keskittää piikkejä sitoutumismotiivien läheisyyteen. Testauksen perusteella ConsensusSummitosoittautui hyödylliseksi menetelmäksi, jolla voi löytää uudenlaista tietoa ChIP-seq-piikkien sijoittumisesta genomiin ja menetelmän tuloksia voidaan hyödyntää esimerkiksi tutkittaessa sitoutumismotiiveja.

Avainsanat: ChIP-seq, Ydinestimointi, Bioinformatiikka, Piikkien koostaminen, ConsensusSummit

ACM-luokat (ACM Computing Classification System, 1998 version):
J.3 LIFE AND MEDICAL SCIENCES–Biology and genetics

UNIVERSITY OF EASTERN FINLAND, Faculty of Science and Forestry, Kuopio
School of Computing
Computer Science

Student, Pauli Tuoresmäki: Aggregation of ChIP-seq peaks using kernel density estimation
Master's Thesis, 59 p., 1 appendix (1 p.)
Supervisors of the Master's Thesis: PhD Wilhelmiina Hämäläinen and PhD, Adj. Prof. Sami Heikkinen
November 2015

Abstract: The focus of biomedical research has been shifting from the laboratory to data analysis. One reason for the shift are methods like ChIP-seq which produce a lot of data with less labor and moderate cost. ChIP-seq is used to study the binding of proteins to DNA. One ChIP-seq experiment produces millions of sequence reads encoded as character strings. Data from ChIP-seq experiments that were used in published research are usually required to be uploaded to public databases. This has led to an increase in available data which in turn has increased the demand for methods that can integrate data from many experiments. To identify the binding sites of the protein the data needs to be processed to ChIP-seq peaks. ChIP-seq peaks from one ChIP-seq experiment can be easily used in research but using data from more than one experiment can produce more information. However, the integration of more than a few experiments has proved to be challenging. That is why methods have concentrated on pairwise comparisons and methods that integrate or aggregate data have not been developed. The purpose of this study was to investigate how ChIP-seq peaks could be aggregated and whether kernel density estimation could be used for it. Based on this investigation we develeped a new method, ConsensusSummit, for aggregation of ChIP-seq-peaks based on kernel density estimation. Kernel density estimation is a method for estimating density functions and it allows to use peak density information as a basis for aggregation. Using data from many ChIP-seq experiments the ConsensusSummit-method aggregates close-by peaks to a "Consensus peak" and leaves lone peaks separate. The method was tested on publicly available data which was selected because of existing comparable results. In addition to comparison we searched for suitable parameters for ConsensusSummit. We also studied how well the method can center peaks around binding motifs. Based on the tests ConsensusSummitproved to be a useful method which can give new kind of information about the placing of ChIP-seq-peaks in the genome. The results of the method can, for example, be used for investigating binding motifs.

Keywords: ChIP-seq, Kernel density estimation, Bioinformatics, Peak aggregation, ConsensusSummit

CR Categories (ACM Computing Classification System, 1998 version):
J.3 LIFE AND MEDICAL SCIENCES–Biology and genetics

# Sanasto

Osaan tämän sanaston sanojen määritelmistä on olemassa sekä tarkempia että laajempiakin määrittelyjä ja merkityksiä. Seuraavat määritelmät on yksinkertaistettu tekstin ymmärtämisen kannalta riittävään tarkkuuteen.

| | |
|---|---|
| Alipehmennys | Undersmoothing. Estimoinnissa käytetty termi, jossa datan yksityiskohdat eivät ole vielä tarpeeksi hävinneet, jotta yleisempi rakenne olisi hahmotettavissa. |
| Artefakti | ChIP-seq-menetelmästä puhuttaessa: jäänne jostakin aikaisemman vaiheesta, joka aiheuttaa virhettä |
| ChIP | Kromatiini-immunopresipitaatio, Chromatin Immunoprecipitation. Menetelmä, jolla voidaan vasta-aineen avulla eristää ne osat DNA:ta johon tutkittava proteiini sitoutuu (ks. Luku 2.2.1). |
| ChIP-seq | ChIP-sekvensointi. Geenitutkimuksen menetelmä, jolla tutkitaan proteiinien sitoutumista DNA:han (ks. Luku 2). |
| ChIP-seq-piikki | Genominen sijainti, johon tutkittu transkriptiofaktori on todennäköisesti ollut sitoutuneena. (ks. Luku 2.3.2). |
| ChIP-seq-koe | ChIP-seq-menetelmällä tehty koe, johon kuuluu yhden tutkittavan kohteen kaikki replikaatit. Yhdessä ChIP-seq-tutkimuksessa voidaan tehdä monta ChIP-seq-koetta. |
| D-vitamiinireseptori | Transkriptiofaktori, joka käyttää D-vitamiinia DNA:han sitoutumisensa säätelyyn. |
| DNA | Deoxyribonucleic acid, Deoksiribonukleiinihappo. Nukleotideistä koostuva biomolekyyliketju, joka muodostaa kaikkien solujen geneettisen materiaalin (ks. myös DNA-säie). |
| DNA-kirjasto | Myös sekvenssikirjasto. DNA-materiaalikirjasto, joka tehdään kromatiini-immunopresipitaatiosta saadusta näytteestä sekvensointia varten. |

| | |
|---|---|
| DNA-säie | DNA-koostuu kahdesta eri säikeestä (+ ja -), jotka ovat toistensa peilikuvia. |
| Emäs | Base. Emäksinen nukleotidi, Emäkset muodostavat DNA:n. Vastinemäksensä kanssa saavan aikaan DNA:n kaksisäikeisen rakenteen muodostaen emäspareja. Käytetään DNA:n mittayksikkönä. (Ks. emäspari (bp) ja nukleotidi.) |
| Emästunnistaja | Sekvensoinnissa käytetty ohjelma, joka määrittää sekvensointilaitteen ottamista kuvista DNA:n nukleotidisekvenssin. |
| Entsymaattinen pilkkominen | Entsyymi-proteiinin avulla tapahtuvaa pilkkomista, joka yleensä tapahtuu entsyymille tyypillisestä kohdasta. |
| ep | Base pair (bp), emäspari. DNA:n kaksisäikeisessä rakenteessa molempien säikeiden vastakkaiset emäkset muodostavat yhdessä emäsparin. Käytetään mittayksikkönä DNA:n pituudelle emäs- ja nukleotidi-termien ohessa. (Ks. emäs ja nukleotidi.) |
| Fragmentti | Pala pilkottua DNA:ta, joiden pituus yleensä vaihtelee muutamista emäspareista satoihin tai tuhansiin, pilkkomistavasta riippuen. |
| Genomi | Perimä, Genome. Termi joka kattaa koko kyseessä olevan organismin DNA:n eli geneettisen informaation. |
| Genominlaajuinen | Koko genomin/perimän kattava. Esimerkiksi genominlaajuinen DNA-sekvensointi, jossa sekvensoidaan sekvensoitavan organismin koko perimä samanaikaisesti. |
| Genominen toistojakso | Genomissa olevia joskus isojakin alueita, jotka toistuvat monessa paikkaa genomia. |
| Genomiin rinnastus | Rinnastuksessa verrataan kahden sekvenssin samankaltaisuutta. Genomiin rinnastuksessa toisena sekvenssinä on koko genomi, josta yritetään löytää samanlainen sekvenssi kuin verrattavassa sekvenssilukemassa. |

| | |
|---|---|
| Huippupiste | Piikin sisällä vahvimman signaalin omaava yksittäinen nukleotidi. |
| Kohina | Noise. Signaalien yhteydessä käytetty termi, joka kuvaa satunnaista taustahäiriötä, joka ei riipu tutkittavasta kohteesta. |
| Koostepiikki | Usean eri ChIP-seq-kokeen saman genomisen sijainnin piikeistä muodostettu piikki. (ks. Luku 4) |
| Kromatiini | Koko perimä eli DNA, mutta laajempi termi. Kromatiini kattaa myös DNA:han kuulumattomia osia, kuten DNA:n pakkautumiseen vaikuttavat proteiinit (vrt. genomi). |
| Kromosomi | Genominen DNA jakautuu fyysisesti erillisiin kromosomeihin. Kromosomeja on eri eläimillä eri määrä. Esimerkiksi ihmisellä on 24 eri kromosomia. |
| Lukemasignaali | Sekvenssilukemasignaali. Rinnastetuista sekvenssilukemista muodostettu numeerinen signaali, joka kertoo sekvenssilukemien määrän kussakin sijainnissa. |
| Mitokondriaalinen DNA | Solussa on genomisen DNA:n lisäksi myös mitokondriaalista DNA:ta, joka ei teoriassa kuulu genomiin. |
| Nukleotidi | Nukleiinihappojen (esim. DNA) rakenneosa. Esiintyy myös muualla solussa. DNA:ssa esiintyy neljää eri nukleotidiä, joiden perusteella DNA:n informaatiota luetaan. DNA:n nukleotidejä merkitään jokaista omalla kirjaimellaan (A, C, G, T) sekvenssiä tutkittaessa. Saatetaan joissakin tapauksissa käyttää mittayksikkönä DNA:n pituudelle. |
| PCR | Polymerase Chain Reaction, polymeraasiketjureaktio. Menetelmä jolla geneettistä materiaalia voidaan monistaa. |
| Piikkien haku | Menetelmä, jossa rinnastetusta sekvenssilukemadatasta muodostetaan piikkejä. |

| | |
|---|---|
| Proteiini | Valkuaisaine. Solujen tuottamia isoja ja usein monimutkaisia molekyylejä, jotka ovat elintoiminnoille välttämättömiä. |
| NGS | Next generation sequencing, seuraavan sukupolven sekvensointi. Genominlaajuiseen sekvensointiin perustuvista menetelmistä käytettävä termi. (Esimerkiksi ChIP-seq ja RNA-seq ovat NGS-menetelmiä.) |
| Replikaatti | Näytteestä tai tutkimuksesta tehty toisinto, jonka tarkoituksena on parantaa tulosten luotettavuutta. |
| Rinnastuvuus | Mappability. Kertoo kuinka suurelle osalle sekvensseistä löytyy vastinkohta vain yhdestä paikasta genomissa. |
| Sekvenssi | Jonossa/järjestyksessä oleva asia. Esimerkiksi DNA-sekvenssi koostuu tietyssä järjestyksessä olevista nukleotideistä, jotka voidaan esittää myös merkkijonona (Ks. nukleotidi). |
| Sekvensointi | Nukleiinihappojen (esim. DNA) nukleotidijärjestyksen määrittäminen. |
| Sekvensointisyvyys | Kertoo miten paljon sekvensoinnista halutaan sekvenssilukemia. |
| Sekvenssilukema | Sequence read. Sekvensoinnista saatavan nukleotidisekvenssin nukleotidien järjestys koodattuna merkkijonoksi laatuarvoineen. Esimerkiksi ChIP-seq-menetelmän tuottama raakadata on miljoonia sekvenssilukemia. |
| Sitoutumismotiivi | Jokaiselle transkriptiofaktorille ominainen nukleotidisekvenssi, johon se DNA:ssa ensisijaisesti sitoutuu. |
| Sonikointi | DNA:n pilkkomismenetelmä, jossa ultraäänen avulla pilkotaan DNA sattumanvaraisenkokoisiin fragmentteihin. (ks. fragmentti) |
| Solulinja | Viljeltävissä oleva solutyyppi. Esimerkiksi ihmisen solut eivät normaalisti ole viljeltävissä maljalla, joten pitkäkestoiseen solujen tutkimiseen joudutaan käyttämään solulinjoja. Yleensä solulinjat ovat lähtöisin syöpäsoluista. |

| | |
|---|---|
| Transkriptiofaktori | Proteiinityyppi, joka pystyy sitoutumaan DNA:han. Transkriptiofaktoreita on paljon erilaisia ja yleensä säätelevät geenien ilmentymistä. |
| Templaatti | Sekvensoinnissa se DNA-fragmentti, jonka sekvenssi halutaan selvittää. |
| Vasta-aine | Molekyyli, joka sitoutuu tietynlaisiin proteiinirakenteisiin. Vasta-aineet voivat sitoutua vain tiettyyn proteiiniin (spesifinen) tai useampiin (epäspesifinen). |
| Ydinestimointi | Kernel density estimation, KDE. Tiheysfunktion estimointimenetelmä (Ls. Luku 3). |
| Ydinestimaatti | Ydinestimoinnin tuloksena saatava tiheysjakauma arvio. |
| Yliedustettu | Sekvenssilukeman tapauksessa tarkoittaa lukemia, jotka ovat täsmälleen samasta paikasta genomia useammin kuin sekvensointisyvyyden perusteella on todennäköistä. |
| Ylipehmennys | Oversmoothing. Estimoinnissa käytetty termi, jossa datan kaikki yksityiskohdat ovat hävinneet, jopa yleiset trendit. |

# Sisältö

# 1 Johdanto

Biolääketieteen tutkimuksen painopiste on lisääntyvissä määrin siirtymässä datan käsittelyyn ja analysointiin. Painopisteen muutos johtuu muun muassa automaattisista laboratoriomenetelmistä, joilla voidaan tuottaa suuria määriä dataa hyvin lyhyessä ajassa, esimerkiksi ihmisen ihmisen koko perimästä. Datamäärän kasvu on luonut tarpeen automaattisille analysointimenetelmille.

*ChIP-seq* on geenitutkimuksen menetelmä, jolla voidaan tutkia proteiinin sitoutumista DNA:han. Yhdellä *ChIP-seq-kokeella* saadaan miljoonia nk. *sekvenssilukemia*. Sekvenssilukemat ovat merkkijonomuotoinen esitys DNA-sekvenssistä, johon kiinnostuksen kohteena oleva proteiini sitoutuu tutkitussa näytteessä. Sekvenssilukemat voidaan jalostaa *ChIP-seq-piikeiksi*, jotka kertovat proteiinin sitoutumispaikan sijainnin tutkittavan organismin perimässä sekä sitoutumisen voimakkuuden.

Keskeinen ongelma ChIP-seq-piikkien tutkimuksessa on, miten tunnistaa aidot ja biologisesti merkitsevät piikit kaikkien piikkien joukosta. Biologisen merkitsevyyden määrittelyssä auttaisi, jos voitaisiin yhdistää tai vertailla tuloksia useammasta eri ChIP-seq-tutkimuksesta. Saatavilla olevan datan määrän yhä kasvaessa useiden eri ChIP-seq-tutkimusten kokeiden hyödyntäminen uusissa tutkimuksissa muuttuu koko ajan oleellisemmaksi uuden tiedon löytämisessä. Usean eri ChIP-seq-kokeen piikkien samanaikainen hyödyntäminen on kuitenkin osoittautunut haasteelliseksi.

Yksi vaihtoehto usean eri ChIP-seq-kokeen hyödyntämiseksi on ChIP-seq-piikkien koostaminen. Koostamisella tarkoitetaan tässä yhteydessä usean eri ChIP-seq-kokeen yhdistämistä niin että samassa genomisessa sijainnissa useassa eri kokeessa olevat ChIP-seq-piikit on esitetty koosteessa vain yhdellä yhteisellä *koostepiikillä*. ChIP-seq-piikkien koostamiseen ei juurikaan ole aiempia ratkaisuja, lukuunottamatta muutamia menetelmiä, joiden pääpaino on enemmän piikkien vertailussa kuin yhdistämisessä. Tämän tutkimuksen tavoitteena oli selvittää, miten ChIP-seq-piikkejä voidaan koostaa sekä voisiko *ydinestimointia* soveltaa piikkien koostamisessa?

ChIP-seq:stä ja ydinestimoinnista löytyy paljon kirjallisuutta, josta etsittiin aiempia tutkimustuloksia ChIP-seq-piikkien koostamisesta sekä tietoa ydinestimoinnista ja sen soveltamisesta ChIP-seq-piikeille. Ydinestimointiin perustuvan koostamismenetelmän puutteen vuoksi kehitettiin uusi ydinestimointiin perustuva ChIP-seq-piikkien koostamismenetelmä. Uuden menetelmän toimintaa testattiin julkisella datalla tarkastelemal-

la menetelmän tuottamia koostepiikkejä sekä vertailemalla koostepiikkejä vastaavanlaiseen dataan.

Tämän tutkimuksen tärkein osa on uusi ChIP-seq-piikkien koostamiseen tarkoitettu menetelmä, ConsensusSummit, ja sen testaus. Muita vastaavanlaisia menetelmiä ei tiettävästi ole kehitetty. Menetelmää käytettiin jo kehitysvaiheessa D-vitamiinireseptorin sitoutumisen tarkasteluun ihmisen genomissa julkaisussa

Pauli Tuoresmäki, Sami Väisänen, Antonio Neme, Sami Heikkinen, ja Carsten Carlberg. Patterns of genome-wide VDR locations. *PloS ONE*, 9(4):e96105, 2014. doi: 10.1371/journal.pone.0096105

Kyseisessä tutkimuksessa ConsensusSummit-menetelmän tuottaman koosteen avulla tarkasteltiin D-vitamiinireseptorin sitoutumisen eroja ihmisen eri kudossoluissa sekä muodostettiin kokonaiskuva D-vitamiinireseptorin sitoutumisesta ihmisen koko perimässä.

Tämän tutkielman rakenne on seuraava: Luvussa 2 esitellään ChIP-seq-menetelmä, ChIP-seq:n tuottama data ja miten datasta muodostetaan ChIP-seq-piikkejä. Luvussa 3 käsitellään puolestaan ydinestimointia ja sen sovelluksia ChIP-seq:ssä. Luvussa 4 esitellään uusi ConsensusSummit-menetelmä ja tarkastellaan sen suhdetta aiempiin koostamis- ja vertailumenetelmiin. Luvussa 5 raportoidaan ConsensusSummit-menetelmän empiirisen testauksen tulokset. Johtopäätökset on esitetty luvussa 6

# 2 ChIP-seq-data ja sen analyysi

ChIP-seq-menetelmä ja sillä tuotettu data ovat keskeisessä roolissa tässä tutkimuksessa, joten menetelmän läpikäynti ja avaaminen on tärkeää. Tässä luvussa käydään tarkemmin läpi millainen ChIP-seq-menetelmä on, minkälaista dataa sillä saadaan sekä miten saatua dataa käsitellään ja tulkitaan. Luvussa käydään myös läpi ChIP-seq:n virheenlähteitä ja ongelmia, joiden tietämisestä on hyötyä tulosten tulkinnassa.

## 2.1 ChIP-seq-menetelmän yleiskuvaus ja kehitys

*ChIP-seq* tai ChIP-sekvensointi on geenitutkimuksen menetelmä, jolla tutkitaan esimerkiksi kiinnostuksen kohteena olevan proteiinin sitoutumista DNA:han. Tässä opinnäytteessä keskitytään vain DNA:han sitoutuvien proteiinien eli *transkriptiofaktoreiden* ChIP-seq-tutkimukseen, mutta esimerkiksi Farnham (2009) ja Park (2009) ovat käsitelleet ChIP-seq-menetelmän muitakin käyttötarkoituksia katsausartikkeleissaan. ChIP-seq-menetelmän vahvuus on sen antama genominlaajuinen tilannekuvaus transkriptiofaktorin sitoutumisesta DNA:han. Kuvaus muodostetaan miljoonista ympäristön ja genetiikan vaikutuksille alttiina olevista soluista (Furey, 2012). ChIP-seq-menetelmä ja sen tuottama raakadata kuvataan tarkemmin luvussa 2.2. Menetelmän teknisyyden vuoksi Liu ym. (2010) on koonnut artikkelin vastauksena ChIP-seq-menetelmän usein herättämiin kysymyksiin.

ChIP-sekvensointi on vielä nuori menetelmä, joka oli ensimmäisiä *seuraavan sukupolven sekvensointimenetelmien* (NGS, next generation sequencing) sovelluksia (Park, 2009). Menetelmää käytettiin ensimmäisen kerran vuonna 2007 (esim. Johnson ym., 2007). ChIP-seq-menetelmän avulla on tehty tieteellisiä läpimurtoja, esimerkiksi transkriptiofaktoreiden tärkeydestä taudeissa, sekä kumottu että vahvistettu väittämiä transkriptiofaktoreiden luokittelusta ja toiminnasta (Farnham, 2009). ChIP-seq-menetelmää tukemaan on kehitetty muitakin genominlaajuisiin sekvensointeihin perustuvia tekniikoita erilaisiin käyttötarkoituksiin (ks. Furey, 2012), mutta niihin ei keskitytä tässä.

ChIP-sekvensointia käytetään paljon bio-alojen tutkimuksissa sen antaman suuren datamäärän ja siihen suhteutetun edullisen hinnan vuoksi. Tutkimuksissa tuotettu raakadata on yleensä saatavilla isoissa ja hyvin ylläpidetyissä tietokannoissa (ks. Liite 1). Isot tietokannat on perustettu takaamaan datalle mahdollisimman pitkäaikai-

nen saatavuus, etenkin kyseiseen dataan perustuvien julkaisujen yhteydessä. Käytetyn raakadatan pitkäaikainen julkinen saatavuus on ollut positiivinen seuraus ChIP-seq-menetelmän virheenlähteistä ja ongelmista.

Vaikka dataa on paljon, ei pelkästä ChIP-seq-raakadatasta pystytä päättelemään kovin paljoa ilman pientä käsittelyä tai analysointia erilaisilla ohjelmistotyökaluilla. Oletuksena raakadatasta on tarkoituksena jalostaa ChIP-seq-piikkejä, jotka kertovat mihin kyseessä oleva proteiini sitoutuu kohdesolujen DNA:ssa. Raakadatan käsittelyyn on vaikeaa määritellä yleispäteviä käytäntöjä, koska ChIP-seq-menetelmässä on monia muuttujia: esimerkiksi solutyyppi, solun olotila sekä käytetty transkriptiofaktori. Joitakin yleisiä toimintaohjeita datankäsittelyyn on kuitenkin julkaistu, jotta raakadatan käsittelystä saataisiin laadukkaampaa ja yhdenmukaisempaa (Bailey ym., 2013; Landt ym., 2012). Raakadatan jalostusta käsitellään tarkemmin luvussa 2.3.

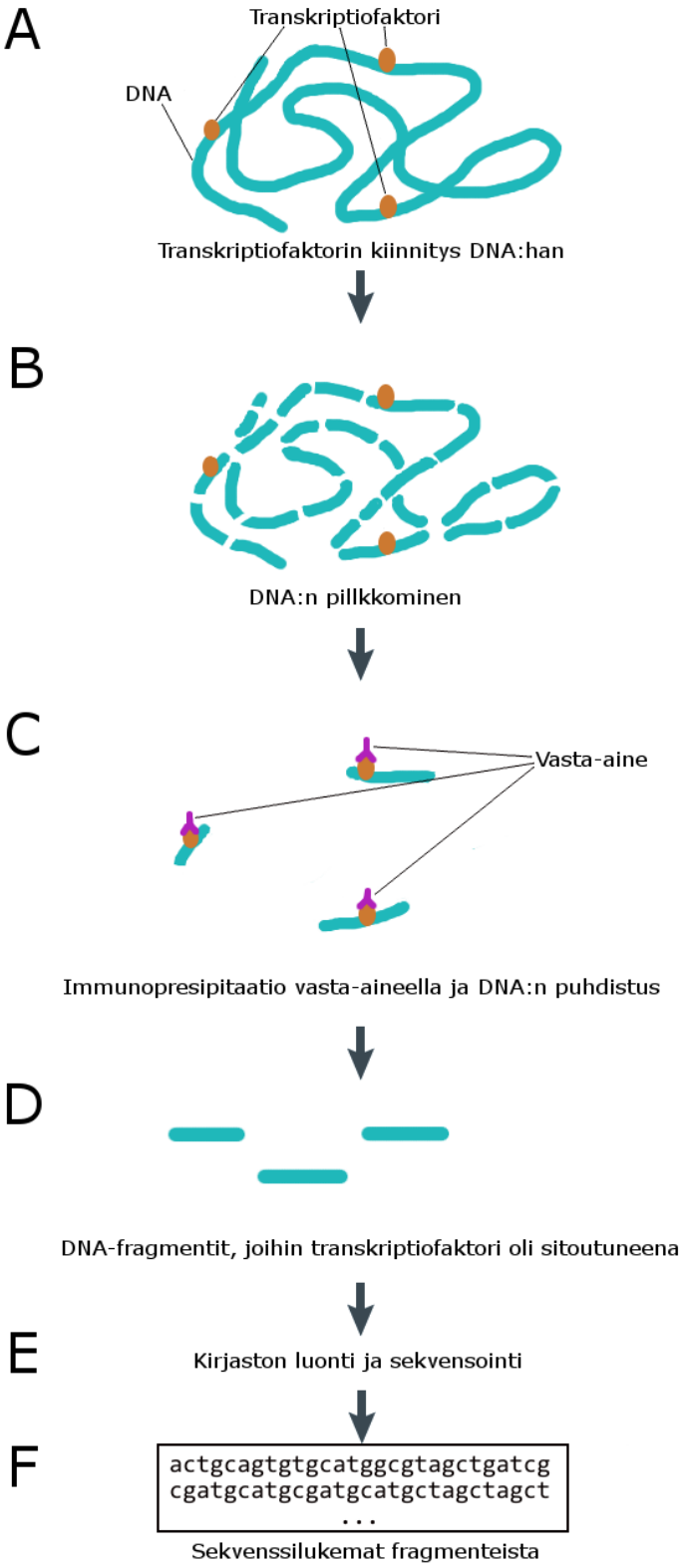
Paljon dataa sisältävistä julkisista tietokannoista johtuen olisi kysyntää menetelmille, joilla eri ChIP-seq-kokeista saatua dataa voisi yhdistää ja vertailla. Monen tutkimuksen datojen yhdistämisen kautta voisi olla mahdollista saada uutta tietoa esimerkiksi geenien säätelystä ja niihin vaikuttavista tekijöistä.

## 2.2 Datan tuottaminen

ChIP-seq-menetelmän tarkoituksena on tuottaa tutkittavista soluista merkkijonomuotoista DNA-sekvenssidataa alueista, joihin kiinnostuksen kohteena oleva proteiini sitoutuu. Käytännössä ChIP-seq yhdistää kaksi menetelmää, kromatiini-immunopresipitaation ja genominlaajuisen DNA-sekvensoinnin, jotka esitellään seuraavissa aliluvuissa. Koko ChIP-seq-menetelmän päävaiheet on esitetty Kuvassa 1.

### 2.2.1 Kromatiini-immunopresipitaatio (ChIP)

*Kromatiini-immunopresipitaation* (Chromatin Immunoprecipitation, ChIP) tarkoituksena on rikastaa DNA:n osat, joihin kiinnostuksen kohteena oleva proteiini sitoutuu. ChIP on suorin tapa tunnistaa yksittäisten DNA:han sitoutuvien proteiinien sitoutumispaikka (Furey, 2012). Tyypillinen kromatiini-immunopresipitaatio tarvitsee vähintään $10^7$ samankaltaista solua ja siitä saa parhaimmillaan nanogrammoja DNA:ta (Park, 2009). Uudemmilla menetelmillä pyritään pienentämään tarvittavien solujen määrää ja




Kuva 1: ChIP-seq-datan tuottamisen vaiheet.

vähentämään DNA:n tarvetta (Furey, 2012).

Kromatiini-immunopresipitaatiossa tutkimuksen kohteena olevia soluja käsitellään formaldehydillä, joka stabiloi kaikki proteiini-DNA-sidokset mukaan lukien kiinnostuksen kohteena olevan DNA:han sitoutuvan proteiinin (Kuva 1A). Sitoutumisen vahvistamisen jälkeen soluista erotetaan niiden perintöaines eli *kromatiini*, joka pilkotaan *entsymaattisesti* tai *sonikoimalla* eli ultraääntä käyttäen (Kuva 1B). Sonikoinnilla pyritään tuottamaan lyhyitä, noin 200 - 600 emäsparin pituisia *DNA-fragmentteja* eli DNA:n palasia. Seuraavaksi pilkottu kromatiini immunopresipioidaan, joka tarkoittaa tutkittavan proteiini-DNA-kompleksin erottamista muusta kromatiinimateriaalista spesifisen vasta-aineen avulla (Kuva 1C). Puhdistuksen jälkeen eroteltu DNA-proteiinikompleksi rikotaan ja komplekseista saatu DNA (Kuva 1D) tutkitaan halutulla tavalla, joka esimerkiksi ChIP-seq:n tapauksessa on sekvensointi. (Park, 2009)

Kromatiini-immunopresipitaation käyttäminen edellyttää aiempaa tietoa tutkimuksen kohteena olevasta proteiinista, koska kyseiselle proteiinille täytyy olla saatavilla *spesifinen vasta-aine* (Furey, 2012). Spesifinen vasta-aine sitoutuu tehokkaasti vain tutkimuksen kohteena olevaan proteiiniin ja mahdollisimman vähän muuhun. Vasta-aineen spesifisyys on yksi menetelmän tärkeimmistä vaatimuksista ja siksi vasta-aineen kehitys ja validointi on tärkeää, mutta myös työlästä (Furey, 2012). Spesifisellä vasta-aineella saadaan parempilaatuista dataa vähäisestä määrästäkin DNA:ta, koska muuhun kuin haluttuun proteiiniin sitoutunutta DNA:ta tulee mukana vähän. Vasta-aineita on kaupallisesti saatavilla monille eri proteiineille, mutta niiden laatu vaihtelee vasta-aineesta riippuen ja joskus myös eri valmistuserien välillä. Joissakin testeissä jopa 20 - 35 % vasta-aineista on osoittautunut huonolaatuisiksi (Park, 2009).

Kromatiinin pilkkoutumiseen yleensä vaikuttaa sen rakenne, sillä tiukemmin pakkautuneet osat hajoavat helpommin kuin löyhemmät osat. Tästä johtuen saatetaan fragmentteja saada epätasaisesti joistakin osista kromatiinia (Park, 2009). Fragmenttien pituus aiheuttaa myös resoluutio-ongelmia, koska proteiinit sitoutuvat yleensä vain 6-20 emäsparin pituiselle alueelle (Furey, 2012).

Immunopresipitaatiovaiheessa tulevia virheitä koetetaan eliminoida analyysivaiheessa käyttämällä erikseen käsiteltäviä kontrollinäytteitä, joita on kolmea yleisesti käytettyä tyyppiä. Selvästi yleisintä on käyttää kontrollina pilkottua näyte-DNA:ta, jolle ei tehdä immunopresipitaatiota (input DNA). Toinen tapa on käyttää kontrollina näyte-DNA:ta, jolle on tehty immunopresipitaatio ilman vasta-ainetta. Kolmas tapa on käyttää DNA:ta

immunopresipitaatiosta, joka on tehty epäspesifisellä vasta-aineella proteiinille, jonka ei pitäisi sitoutua DNA:han. Input-DNA on kuitenkin selvästi käytetyin ja se korjaa kromatiinin pakkautumisesta ja pilkkoutumisesta johtuvia vaihteluita. (Park, 2009)

### 2.2.2 Sekvensointi (seq)

*Sekvensoinnin* (sequencing, seq) tarkoituksena on muodostaa halutusta DNA:sta sitä vastaava tietokoneella käsiteltävä merkkijono. Genominlaajuista sekvensointia varten ChIP:stä saadusta DNA:sta täytyy valmistaa *DNA-kirjasto*, jota tehdessä valikoidaan tietyn pituiset (yleensä noin 150 - 300 emäsparia pitkät) fragmentit (Park, 2009). Valmistettu DNA-kirjasto sekvensoidaan NGS-sekvensointilaitteella (Kuva 1E-F). Sekvensointiin on olemassa muutama laitekohtainen, hieman toisistaan poikkeava menetelmä (ks. Metzker, 2010). Perusidea laitteissa on kuitenkin sama.

Yksinkertaistettuna sekvensoinnissa kirjaston DNA:ta monistetaan *templaateiksi*, jonka jälkeen templaattien sekaan lisätään yksitellen värjättyjä *emäksiä*. Emäkset sitoutuvat vastinpariinsa templaatissa. Jokaisen emäksen lisäyksen jälkeen otetaan korkearesoluutioinen kuva, jossa emäksen lisäyksestä johtuvat värinmuutokset näkyvät. Kuvatiedostot muunnetaan *emästunnistajalla* (base caller) sekvensseiksi. Lopputuloksena saadaan tiedosto halutun pituisia merkkijonoiksi koodattuja DNA-sekvenssejä eli *sekvenssilukemia* (sequence read). Saatu sekvenssilukema ei siis ole koko alkuperäisen DNA-fragmentin pituinen vaan tietyn mittainen osa fragmentin alkupäästä. Näin kaikki lukemat ovat samanpituisia riippumatta fragmenttien pituuksista. Joillakin laitteilla DNA-sekvenssien lisäksi saadaan myös kullekin sekvenssin nukleotidille laatuarvot, jotka kertovat millä varmuudella nukleotidi on määritetty oikein. Mahdollisia laatuarvoja voidaan käyttää hyödyksi myöhemmin dataa käsiteltäessä. (Metzker, 2010). Sekvensointivirheitä tapahtuu nykyisin harvoin (Park, 2009), mutta sekvensointilaitteissa on eroja. Esimerkiksi Illuminan sekvensointilaite tuottaa huonolaatuisia nukleotidejä sekvenssilukeman loppupäähän (Furey, 2012).

Sekvenssilukemat esitetään usein *fastq-formaatissa* (Cock ym., 2010). Kuvassa 2 on annettu näyte fastq-formaatissa olevasta sekvenssilukematiedostosta. Ensimmäinen rivi on sekvenssin ID ja siinä on yleensä myös muuta tietoa sekvenssistä. Tässä tapauksessa se sisältää esimerkiksi tietoa sekvensointilaitteesta ja sekvenssin pituuden. Toinen rivi on itse sekvenssi. Kolmas rivi alkaa +-merkillä ja voi sisältää saman tiedon kuin ensimmäinenkin rivi. Neljäs rivi kertoo laatuarvon jokaiselle nukleotidille

sekvenssissä. Laatuarvoina käytetään ASCII-merkkejä, joista jokainen vastaa tiettyä lukuarvoa. Koodeina käytetyt lukuarvot ja niitä vastaavat ASCII-merkit ovat ehtineet vaihdella sekvensointimenetelmien elinkaarien aikana.

Sekvenssilukemia saadaan sekvensoinnin seurauksena nykyisin tilauksen mukaan, tyypillisesti 30 miljoonaa lukemaa. Lukemien määrän yhteydessä puhutaan usein *sekvensoinnin syvyydestä*, joka kertoo miten paljon lukemia sekvensoinnin halutaan tuottavan. Sopivaa syvyyttä sekvensoinnille on hankala määritellä, mutta jos proteiinilla on monta sitoutumispaikkaa, niin yleensä tarvitaan enemmän lukemia, jotta saadaan samanlainen lukematiheys jokaiselle sitoutumisalueelle. Liian vähäinen lukemien määrä puolestaan vaikeuttaa piikkien tunnistamista.

Perustapauksessa sekvensointi tapahtuu vain templaatin toisesta päästä (ns. single-end sequencing), mutta sekvensointi voidaan suorittaa myös templaatin molemmista päistä (ns. paired-end sequencing) (Fullwood ym., 2009). Jälkimmäisellä tavalla saadaan muiden hyötyjen (ks. Korbel ym., 2007) lisäksi parannettua esimerkiksi sekvenssin genomiin rinnastuksen oikeellisuutta (Li ja Homer, 2010). Käytännössä molemmista päistä tehtävällä sekvensoinnilla saavutetut edut ovat kuitenkin vähäisiä suhteessa sen aiheuttamiin suurempiin kustannuksiin ja työmäärään.

Sekvensoinnista saadun raakadatan laatua voidaan arvioida käyttämällä olemassa olevia laadunvarmennustyökaluja. Esimerkkinä FastQC-laadunvarmennustyökalulla voidaan muun muassa koostaa yhteenveto sekvenssilukemien nukleotidien laatuarvoista ja laskea sekvenssikopioiden lukumääriä. Työkalujen tulosten perusteella voidaan luoda yleiskuva datan käyttökelpoisuudesta ja miettiä mahdollisten lisäkäsittelyjen tarvetta.

## 2.3 Datan analysointi

ChIP-seq-datan analysoinnin päätavoitteena on etsiä *ChIP-seq-piikkejä*, jotka kertovat mihin kyseessä oleva proteiini kohdesolujen DNA:ssa sitoutuu. Tätä ennen sekvenssilukemat täytyy kuitenkin rinnastaa tutkittavan eliön, esimerkiksi ihmisen, genomiin, jotta lukemille saadaan niiden genomiset osoitteet.

```
@TCF7L2_Hct116.749996  SOLEXA1_FC30F7HAAXX:2:13:1648:755  length=32
CATATGACTGAGGCGCCCATGGGGGTGGGGGG
+TCF7L2_Hct116.749996  SOLEXA1_FC30F7HAAXX:2:13:1648:755  length=32
<<<<<<<<<<<<<<<<<<;<<<<<<<;<-<<<
@TCF7L2_Hct116.749997  SOLEXA1_FC30F7HAAXX:2:13:97:1343  length=32
CTTGGGAGGCTGAGGCAGGATAATCGCTTGAA
+TCF7L2_Hct116.749997  SOLEXA1_FC30F7HAAXX:2:13:97:1343  length=32
<<<<<<<<<<<<<<<<<<9<9<<<92<<,6<
@TCF7L2_Hct116.749998  SOLEXA1_FC30F7HAAXX:2:13:1471:523  length=32
CCAGGCCCATCCCAATAAACCCCAGCACTAGG
+TCF7L2_Hct116.749998  SOLEXA1_FC30F7HAAXX:2:13:1471:523  length=32
<<<<<<<<<<<<<<<<<<<<<<<:<<;<<<<<
@TCF7L2_Hct116.749999  SOLEXA1_FC30F7HAAXX:2:13:1313:1185  length=32
CGGGGCCAGTGGGCCGGGGCCTAAAAGGCGGC
+TCF7L2_Hct116.749999  SOLEXA1_FC30F7HAAXX:2:13:1313:1185  length=32
<<<<<<<<<<<<<<<:<<<;;<;;;5<<86<2
@TCF7L2_Hct116.750000  SOLEXA1_FC30F7HAAXX:2:13:27:136  length=32
GACCATCCCAGGACCCTGAGCCCCCAACTCTC
+TCF7L2_Hct116.750000  SOLEXA1_FC30F7HAAXX:2:13:27:136  length=32
<<<<<<<<<8<<6<<<<<6<<<<<<(4<2<2<
```

Kuva 2: Esimerkkiote sekvenssilukemia sisältävästä tiedostosta fastq-formaatissa. Otteessa on viiden sekvenssilukeman tiedot, joista jokaisessa on 4 riviä.

### 2.3.1 Genomiin rinnastus

*Genomiin rinnastuksen* (Genome alignment) tavoitteena on löytää jokaisen sekvenssilukeman alkuperäinen sijainti tutkittavan eliön *genomisessa DNA:ssa*. Näin sekvenssilukemiin saadaan liitettyä niiden genomiset koordinaatit eli missä *kromosomissa* ja missä kohden kromosomia ne sijaitsevat. Kaikista seuraavaan sukupolven sekvensointimenetelmistä, myös ChIP-sekvensoinnista, tulevat datat täytyy rinnastaa genomiin. Rinnastukseen on kehitetty useita eri työkaluja, jotka käyttävät eri algoritmeja. Osa rinnastustyökaluista soveltuu paremmin ChIP-seq:lle ja osa muille NGS-menetelmille. Li ja Homer (2010) ovat tehneet kattavan vertailun eri rinnastustyökaluista algoritmeineen. Rinnastusalgoritmit ovat tasapainoilua tarkkuuden, nopeuden, muistinkäytön ja joustavuuden suhteen, joista eri rinnastusalgoritmit painottavat eri asioita (Park, 2009). ChIP-seq-datalle sopii esimerkiksi Bowtie-työkalu (Langmead ym., 2009). Bowtie käyttää indeksoitua genomia saavuttaakseen nopean rinnastuksen ja se sallii myös pieniä määriä erilaisia yhteensopimattomuuksia sekvenssilukeman ja genomin sekvenssin välillä.

Liian lyhyet sekvenssilukemat saattavat tuottaa ongelmia rinnastusvaiheessa. On tutkittu, että 25 nukleotidin pituisista sekvenssilukemista vain noin 80 % on ainutlaatuisia (ts. kyseinen sekvenssi esiintyy genomissa vain yhden kerran), kun taas 43 nukleotidin pituisista sekvenssilukemista ainutlaatuisten osuus on jo 90 % (Whiteford ym., 2005). Mitä enemmän ainutlaatuisia sekvenssilukemia saadaan, sitä vähemmän tulee useaan sijaintiin rinnastuvia sekvenssejä ja *yliedustettuja* sekvenssejä. Yliedustetuilla sekvenssilukemilla tarkoitetaan lukemia, jotka alkavat täsmälleen samasta kohtaa genomia. Ainutlaatuisten lukemien osuutta kaikista lukemista kutsutaan joskus *rinnastettavuudeksi* (mappability). Yksi syy huonoon rinnastuvuuteen ovat lyhyet genomiset toistojaksot, joita on erityisesti nisäkkäillä. Esimerkiksi ihmisen DNA:sta 52 % on toistuvajaksoista. Lyhyet sekvenssilukemat rinnastuvat siis suuremmalla todennäköisyydellä toistojaksoihin ja tätä myötä useaan paikkaan genomissa (Park, 2009).

Ongelmallisista sekvenssilukemista rinnastusvaiheessa pyritään yleensä poistamaan genomiin rinnastumattomat sekvenssilukemat. Useaan eri sijaintiin rinnastuville sekvensseille puolestaan arvotaan yksi paikka monesta yhtä todennäköisestä sijainnista. Yliedustetut sekvenssit voidaan poistaa joko rinnastuksen tai piikkien haun yhteydessä. Poisto kuitenkin jätetään yleensä piikkien hakuvaiheeseen. Yliedustetut lukemat poistetaan, koska ei ole varmuutta johtuuko yliedustus biologiasta vai onko kyseessä

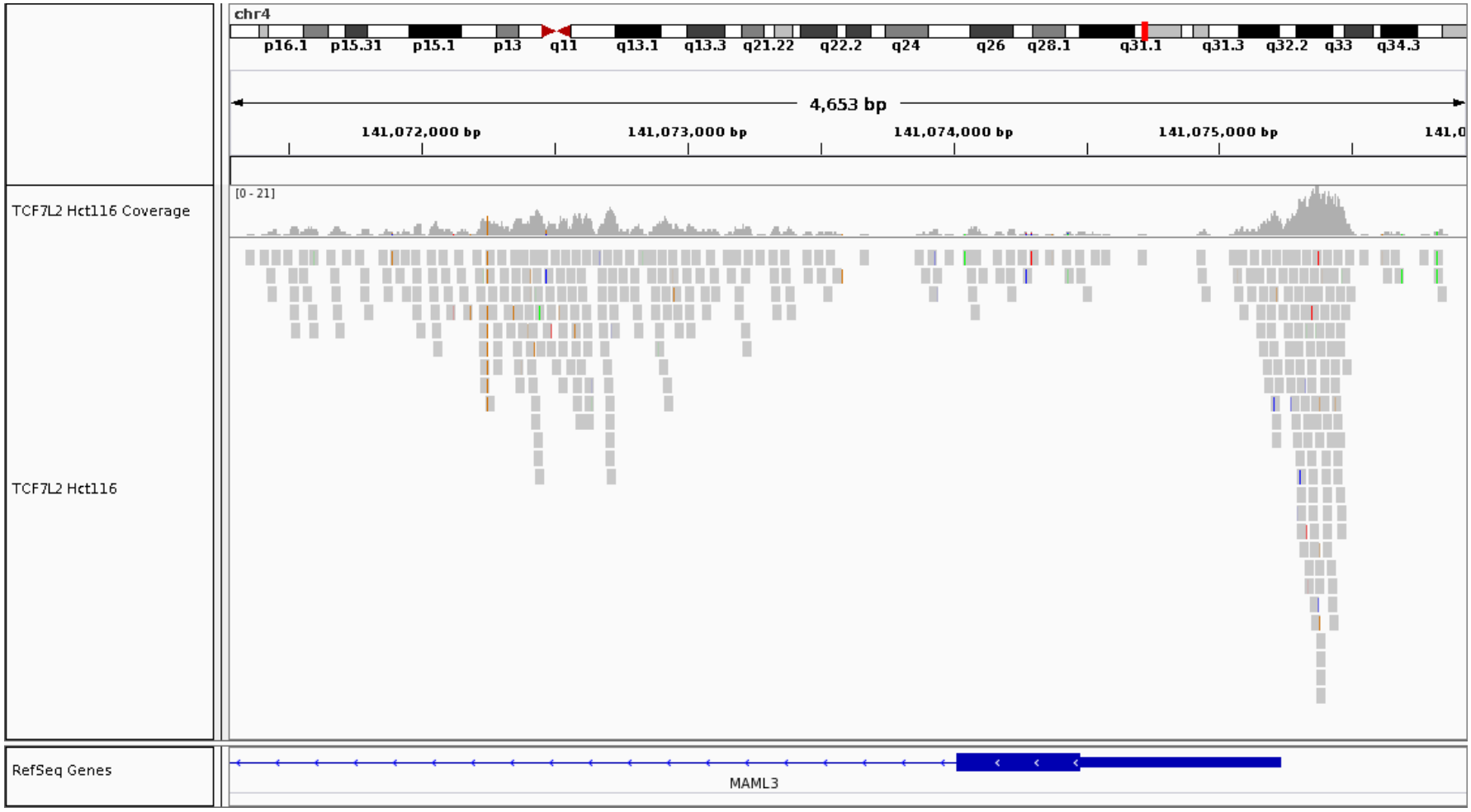
kirjaston luonnissa tapahtunut virhe. Yliedustuksen poisto on siis tasapainoilua todellisten lukemien häviämisen ja menetelmästä johtuvan artefaktin minimoimisen kanssa.

Rinnastuksen jälkeen dataa voidaan ensimmäisen kerran hyvin visualisoida ja silmämääräisesti tutkia. Esimerkki ihmisen genomiin rinnastetusta datasta on Kuvassa 3. Kuvan visualisointi on tehty Interactive Genomics Viewer (IGV) -työkalulla (Robinson ym., 2011). Kuvassa jokainen harmaa suorakulmio on yksi sekvenssilukema, joka on rinnastettu kyseiseen kohtaan genomissa. Kuvan genomisessa sijainnissa on paljon lukemia eli todennäköinen piikki, mutta sen voi varmuudella sanoa vasta piikkien haun jälkeen. Piikkien hakua on käsitelty tarkemmin seuraavassa aliluvussa.

### 2.3.2 ChIP-seq-piikkien haku

*Piikkien haun* tarkoituksena on löytää genomiset alueet, joihin on rinnastunut enemmän sekvenssilukemia kuin puhtaan sattuman kautta olisi odotettavissa (Furey, 2012). Piikkien hakua varten rinnastetusta sekvenssilukemadatasta muodostetaan numeerinen *lukemasignaali*. Lukemasignaali saadaan laskemalla jokaiselle genomiselle sijainnille siihen osuvien sekvenssilukemien määrä. Näin lukemasignaali on vahva alueilla, joissa on paljon lukemia ja heikko alueilla jossa on vähän lukemia. Ympäristöstään merkitsevästi poikkeavat, vahvan lukemasignaalin alueet muodostavat piikkejä, jotka kertovat tutkittavan transkriptifaktorin vahvasta sitoutumisesta kyseiseen sijaintikohtaan. Transkriptiofaktorit sitoutuvat yleensä myös enemmän tai vähemmän sattumanvaraisesti ympäri genomia, mikä aiheuttaa *kohinaa* (noise). Kohinan vaikutusta pyritään minimoimaan, tilastollisilla testeillä sekä vähentämällä näytteen signaalista kontrollinäytteen vastaava signaali.

Piikit luokitellaan yleensä kolmeen eri luokkaan: pistemäisiin, leveisiin ja niiden yhdistelmiin. Transkriptiofaktori-ChIP-seq:ssä pyritään saamaan pistemäisiä piikkejä eli vahvoja mutta kapeita signaaleja. Pistemäisyys on seurausta siitä että transkriptiofaktorit, muutamaa poikkeusta lukuunottamatta, sitoutuvat lyhyeen 6-20 emäsparin mittaiseen DNA-jaksoon (Furey, 2012). Tämä lyhyt DNA-jakso on jokaiselle transkriptiofaktorille omanlaisensa ja sitä kutsutaan *sitoutumismotiiviksi* (binding motif). Sitoutumismotiivin ansiosta lukemien pitäisi kasautua pääasiassa pienelle alueelle kyseisten motiivien läheisyyteen ja näkyä pistemäisinä piikkeinä. Piikeille lasketaan usein *huippupiste* (summit), joka on vahvimman signaalin omaava yksittäinen nukleotidi piikin sisällä. Huippupiste mielletään usein transkriptiofaktorin varsinaiseksi sitoutumispai-



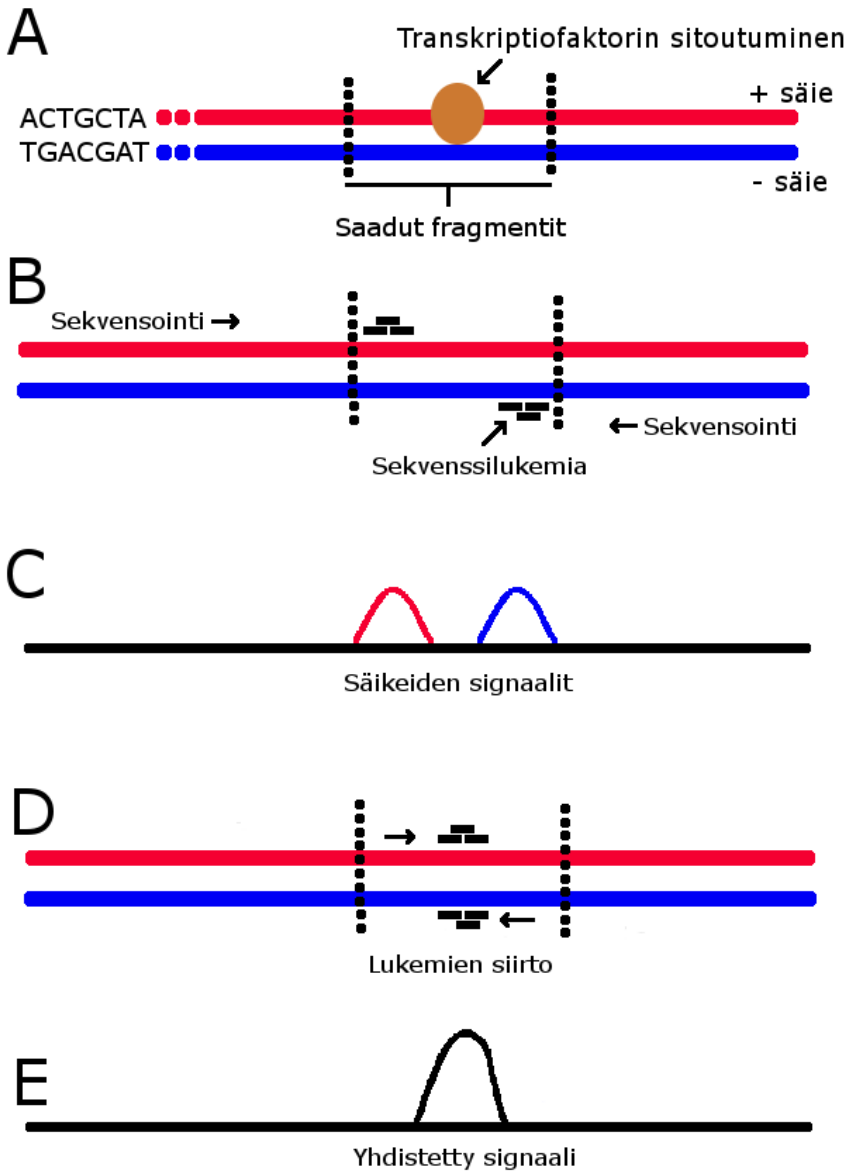

Kuva 3: Visualisointi ihmisen genomiin rinnastetusta ChIP-seq-datasta.

kaksi (Zhang ym., 2008).

Piikkien haussa piikkikandidaatteja karsitaan tutkimalla, onko kandidaattialueelle kertyneiden lukemien lukumäärä merkitsevästi suurempi kuin puhtaan sattuman perusteella olisi odotettavissa. Käytännössä suoritetaan tilastollinen hypoteesin testaus, jonka tuloksena saatava $p$-arvo heijastaa sattuman todennäköisyyttä. Useiden piikkikandidaattien tilastollinen testaus johtaa kuitenkin ns. *moninkertaisen testauksen ongelmaan* (multiple testing -problem, ks. esim. Shaffer, 1995), jolloin epäaitojen, mutta merkitsevyystestin läpäisseiden piikkien lukumäärä kasvaa. Ongelmaan on esitetty useita, mutta ei yhtään universaalisti tyydyttävää ratkaisua. Ongelman vuoksi piikkien hakuun on useita eri työkaluja algoritmeineen, joista osa on keskittynyt tietynlaisiin piikkeihin. Eri työkalujen välillä ei kuitenkaan ole huomattavia eroja (Wilbanks ja Facciotti, 2010). Park:n (2009) mukaan hyvä työkalu ottaa huomioon säiekohtaiset vaihtelut, kontrollinäytteen ja mahdollisesti myös rinnastettavuuden. Yksi esimerkki piikinhakutyökaluista on aiemmin pelkästään kapeisiin piikkeihin erikoistunut MACS (Zhang ym., 2008), joka ottaa huomioon kontrollinäytteet ja säiekohtaiset vaihtelut.

Säiekohtaiset vaihtelut johtuvat DNA:n kaksisäikeisyydestä, jossa toinen säie on toisen peilikuva. Kun DNA:ta ChIP-seq:ssä sekvensoidaan, tulee sekvenssilukema molemmista säikeistä. Koska sekvensointi tapahtuu vain toisesta päästä ja se on vain osa alkuperäistä fragmenttia, tulee eri säikeiden välisille lukemille yleensä fragmentin pituuden verran väliä toisiinsa. Piikkiin pitää yhdistää signaali molemmista säikeistä. Yhdistys tehdään yleensä joko pidentämällä molempien säikeiden lukemia fragmenttien keskipituudella tai siirtämällä lukemia eteenpäin puolella fragmentin keskipituudesta. Näitä säiekohtaisia vaihteluita ja siirtoratkaisua on esitetty Kuvassa 4. (Pepke ym., 2009)

Mahdollisen kontrollinäytteen huomioon ottaminen on tärkeässä roolissa piikkien haussa, koska sekvenssilukemien jakauma ei ole täysin satunnainen. Tämä johtuu siitä että proteiinien taustasitouminen ei ole täysin satunnaista, mikä puolestaan johtuu esimerkiksi aiemmin ChIP:n yhteydessä mainitusta kromatiinin pakkautumisesta. Pelkällä satunnaistaustalla, ilman kontrollinäytettä, laskettuja piikin lukuarvoja, kuten $p$-arvoa, ei pidetä luotettavina. Joissakin tapauksissa pelkän satunnaistaustan käyttö saattaa johtaa jopa kertaluokan kokoisiin virheisiin (Kharchenko ym., 2008). Kontrollinäytteen avulla saadaan siis varmemmin poistettua kohinaa ja laskettua tarkempi signaalin rikastumisen suhde taustaan verrattuna (fold enrichment).




Kuva 4: Periaatekuva sekvenssilukemien säiekohtaisesta vaihtelusta (A-C) ja sen ottamisesta huomioon piikkien haussa (D-E).

Eri piikkien hakualgoritmeille on tehty vertailuja, mutta ne ovat osoittautuneet haastaviksi. Piikkien hakutyökalun suorituskykyä mitataan yleensä kahdella eri tavalla: joko lasketaan saatujen piikkien etäisyyshajonta lähimpään sitoutumismotiiviin tai validoidaan osa piikkialueista käyttäen apuna *polymeraasiketjureaktiota* (Polymerase chain reaction, PCR) (Park, 2009). PCR:llä validointi on kuitenkin nykyisin harvinaistunut. Pienten parannusten tehokkuutta piikkien hakualgoritmeissa on hankala verrata, koska eri datakokoelmille on vain vähän varmennettuja sitoutumispaikkoja. Tämän vuoksi

esimerkiksi Wilbanks ja Facciotti (2010) ovat esittäneet, että paras tapa vähentää väärien piikkien ja kohinan määrää on parantaa tutkimusasettelua ja näytteiden käsittelyä sekä suurentaa biologisten replikaattien määrää algoritmien kehittelyn sijasta.

### 2.3.3 Jatkoanalyysit ja tulosten tulkinta

Piikkihaun jälkeinen analysointi vaihtelee suuresti. Jatkoanalyysi riippuu siitä, mitä tutkittavien transkriptiofaktoreiden sitoutumisella halutaan saada selville. On kuitenkin asioita, jotka vaikuttavat siihen mitä ChIP-seq-piikeistä:stä voidaan ja mitä ei voida tutkia.

Tulosten tulkinnassa kannattaa ottaa huomioon tarvittavien solujen suuri määrä. Koska soluja on paljon, piikit kuvaavat transkriptiofaktorin keskiarvoista sitoutumista solupopulaatiossa. Yksittäisten solujen välillä tiedetään olevan vaihtelua, mutta yksittäisten solujen piikkijakaumia ei pystytä ChIP-seq:ä käyttämällä erottamaan (Furey, 2012). Pieni piikki saattaa siis kuvastaa joko vahvaa sitoutumista pienessä osassa soluja tai heikkoa sitoutumista kaikissa soluissa (Farnham, 2009). Eri solutyyppien, esimerkiksi maksa- ja suolistosolujen, välisiä sitoutumiseroja voidaan siis tutkia vain solupopulaatioita käyttäen. Yksittäisiä soluja verratessa solutyyppien väliset sitoutumiserot ovat useimmiten selvempiä kuin samantyyppisiä soluja verrattaessa (Furey, 2012).

ChIP-seq:llä ei myöskään voida tutkia kuin yhtä transkriptiofaktoria kerrallaan. Tuloksista ei siis voida päätellä onko kyseessä pelkästään tutkittavan transkriptiofaktorin sitoutuminen eli *suora sitoutuminen* vai muiden proteiinien avulla tai niiden kanssa tapahtuva sitoutuminen eli *epäsuora sitoutuminen*. Tulokset eivät myöskään paljasta sitoutumisen vakautta eli miten helposti sitoutuminen tapahtuu tai purkautuu. Tuloksista ei voida myöskään suoraan tulkita sitoutumisen tarkoitusta. (Furey, 2012)

ChIP-seq soveltuu hyvin transkriptiofaktoreiden sitoutumismotiivien tutkimiseen. Joillekin transkriptiofaktoreille on saatu sitoutumismotiivi selville tutkimalla pieni määrä kyseisen transkriptiofaktorin tunnettuja sitoutumispaikkoja ja katsomalla mistä nukleotideistä ne muodostuvat (Farnham, 2009). ChIP-seq:n avulla pystytään tutkimaan suurempi määrä sitoutumispaikkoja ja mahdollisesti parantamaan tunnettujen sitoutumismotiivin oikeellisuutta. On huomattu, että transkriptiofaktorit sietävät jonkin verran vaihtelua sitoutumismotiivissaan. Tämän vuoksi kunkin transkriptiofaktorin sitoutumismotiivi esitetään yleensä matriisina, jossa on annettu kunkin emäksen esiintymis-

tiheys jokaisessa sitoutumismotiivin kohdassa (Furey, 2012). Esimerkiksi Kuvassa 5 on esitetty TCF7L2-transkriptiofaktorin sitoutumismotiivi sekä matriisina että kuvana. Mitä isompi kirjain kuvamuodossa on muihin verrattuna, sitä yleisempi ja siihen perustuen tärkeämpi kyseinen nukleotidi on transkriptiofaktorin sitoutumisessa.

Esiintymistiheysmatriisin avulla voidaan annetulle sekvenssille laskea todennäköisyys sille, sitoutuuko transkriptiofaktori siihen. Isoa sitoutumistodennäköisyyttä pidetään yleisesti mittarina sitoutumisen *voimakkuudesta* (affinity) ja tällaisten alueiden oletetaan tuottavan vahvempia signaaleja (Furey, 2012). Sitoutumismotiiveja on kuitenkin, etenkin nisäkkäiden genomissa, yleensä paljon enemmän kuin löydettyjä sitoutumispaikkoja (Farnham, 2009). Tämä voi johtua siitä että vain tietyt osat genomista ovat soluissa kulloinkin käytössä. On myös alueita, joissa voi olla vahva ja kapea signaali ilman sitoutumismotiiviakin, mutta sitoutumisen syitä ei tiedetä. Solujen määrän vuoksi kyseessä tuskin kuitenkaan on vain sattumanvarainen sitoutuminen (Farnham, 2009). Tällaiset poikkeustapaukset voivat joskus olla hyvinkin mielenkiintoisia tutkittavia.

| | A | C | G | T |
|---|---|---|---|---|
| 1 | 0.832 | 0.013 | 0.097 | 0.058 |
| 2 | 0.005 | 0.501 | 0.485 | 0.009 |
| 3 | 0.737 | 0.002 | 0.001 | 0.260 |
| 4 | 0.022 | 0.002 | 0.001 | 0.975 |
| 5 | 0.001 | 0.901 | 0.095 | 0.003 |
| 6 | 0.997 | 0.001 | 0.001 | 0.001 |
| 7 | 0.989 | 0.005 | 0.005 | 0.001 |
| 8 | 0.995 | 0.001 | 0.003 | 0.001 |
| 9 | 0.058 | 0.001 | 0.940 | 0.001 |
| 10 | 0.203 | 0.135 | 0.592 | 0.070 |
| 11 | 0.280 | 0.307 | 0.356 | 0.057 |
| 12 | 0.445 | 0.197 | 0.165 | 0.193 |

Kuva 5: TCF7L2-transkriptiofaktorin 12 emäsparin pituinen sitoutumismotiivi esitettynä kuvana (yllä) ja matriisina (alla).

# 3 Ydinestimointi

Tässä luvussa kuvataan ydinestimoinnin perusperiaatteet sekä ydinestimoinnin parametrien valintamenetelmiä. Luvussa keskitytään vain yksiulotteiseen dataan, koska moniulotteisen datan ydinestimoinnille ei tämän opinnäytetyön puitteissa ole tarvetta. Lopuksi tarkastellaan ydinestimoinnin käyttöä ChIP-seq-analyyseissä.

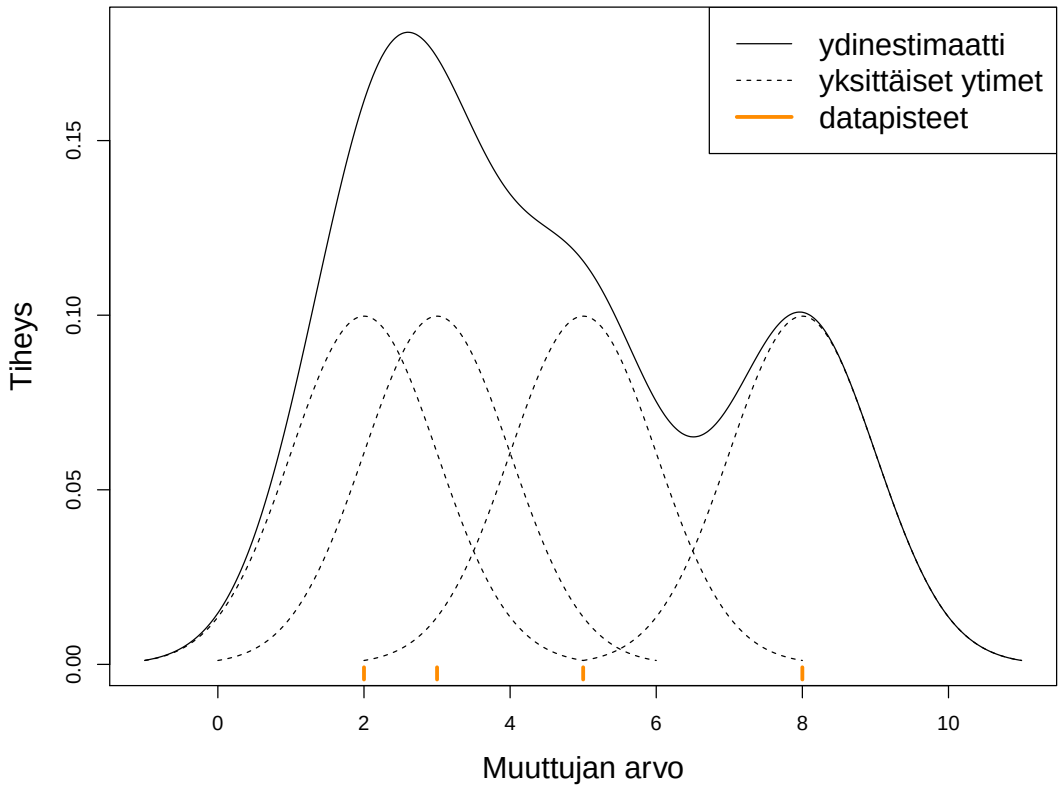
## 3.1 Ydinestimoinnin perusidea

*Ydinestimointi* (Kernel density estimation, KDE, ks. esim. Silverman, 1986) on tiheysfunktion estimointimenetelmä, jolla voidaan kuvata datan jakauma. Tuloksena saatavaa *ydinestimaattia* voidaa pitää histogrammin jatkuvana yleistyksenä. Ydinestimointi on parametriton menetelmä, mikä tarkoittaa sitä ettei tiheysjakauman tarvitse noudattaa jotain tiettyä muotoa, jonka kuvaamiseen riittäisi joukko parametreja. Menetelmän ovat alunperin kehittäneet Rosenblatt (1956) ja Parzen (1962).

Perusideana ydinestimoinnissa jokaisen datapisteen lähiympäristöön sijoitetaan paikallinen tiheysjakauma, joiden yhdistelmänä saadaan koko dataa kuvaava tiheysjakauma. Yksittäisen pisteen vaikutus riippuu paikalliselle tiheysjakaumalle eli ytimelle valitusta muodosta eli *ydinfunktiosta* (kernel function) ja ytimen leveydestä (bandwidth). Ydinestimoinnin toimintaa on havainnollistettu Kuvassa 6. Kuvan tapauksessa ydinestimaatti on laskettu vain neljälle datapisteelle, jotta ydinestimoinnin toimintaa on helpompi havainnollistaa. Kuvassa jokaisen datapisteen (oranssilla) kohdalle on sijoitettu gaussinen ydin, jonka leveys on 1 (katkoviiva). Nämä yksittäiset ytimet yhdistetään ydinestimaatiksi, joka näkyy kuvassa yhtenäisenä viivana.

Histogrammiin verrattuna ydinestimaatti on kehittyneempi tapa arvioida tiheyttä, koska tiheyden kuvaaja on jatkuva ja pehmeä. Ydinestimaatti ei myöskään riipu histogrammin tavoin estimaatin aloituskohdasta tai valittujen luokkavälien leveydestä (Silverman, 1986). Lisäksi ydinestimaatissa on vähemmän harhaa (bias) (Sheather, 2004). Ydinestimointi on kuitenkin laskennallisesti haastavampi. Parametrittomana menetelmänä ydinestimoinnissa ei ole jakaumaoletuksia, mutta ydinestimointia varten joudutaan silti valitsemaan kaksi estimaatin pehmeyteen ja muotoon vaikuttavaa tekijää: ydinfunktio ja ytimen leveys.

Mikäli 1-ulotteinen data koostuu pisteistä $(X_1, ..., X_n)$ saadaan ydinestimaatti $\hat{f}$ pis-




Kuva 6: Ydinestimointia havainnollistava kuva yksinkertaisella fiktiivisellä aineistolla. Ydinestimointiin on käytetty gaussista ydintä ja ytimen leveyttä 1. Datapisteet on asetettu kohtiin 2, 3, 5 ja 8.

teessä $x$ kaavalla

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

missä $K$ on käytetty ydinfunktio ja $h$ on ytimen leveys. Ydinfunktiolta $K$ vaaditaan että se toteuttaa ehdon

$$\int K(x)\, dx = 1,$$

jotta estimaatti on jatkuva (Silverman, 1986; Sheather, 2004). Koska ydinestimaatti saa myös ytimen ominaisuudet, valitsemalla ytimeksi nollan suhteen symmetrisen todennäköisyysjakauman, myös ydinestimaatista tulee todennäköisyysjakauma (Silverman, 1986). Tällaisessa tapauksessa ydin täyttää myös symmetrisyysehdot

$$\int xK(x)\, dx = 0 \quad \text{ja}$$

$$\int x^2 K(x)\, dx = \mu_2(K) > 0,$$

missä $\mu_2(K) > 0$ on funktion $K$ odotusarvo. (Sheather, 2004). Ytimen vaikutus tulokseen on loppujen lopuksi vähäinen, mutta sen valinta vaikuttaa esimerkiksi tilanteissa, jossa estimoitava tiheys ei ole oikeasti symmetrinen (Silverman, 1986). Ytimen leveys puolestaan määrittää miten paljon annettu estimaatti pehmentää/tasoittaa tiheyttä. Ytimen leveys on tärkein tekijä ydinestimoinnissa (Sheather, 2004). Erilaisia ydinfunktioita ja ytimen leveyden määrittäminen on käsitelty erikseen luvuissa 3.2 ja 3.3.

Tavanomaisessa ydinestimoinnissa käytetään samaa ytimen leveyttä koko datassa, mutta on olemassa myös *mukautuvan leveyden ydinestimointi* (adaptive/variable bandwidth kernel density estimation). Mukautuvan leveyden ydinestimoinnissa leveyttä säädellään datan mukaan estimaattia muodostettaessa (ks. esim. Terrell ja Scott, 1992; Sain ja Scott, 1996). Mukautuvuus saadaan aikaiseksi ottamalla huomioon käsittelyssä olevan datapisteen etäisyys valinnaiseen määrään lähimpiä datapisteitä. Estimaatin laskenta on tästä syystä mukautuvalla leveydellä hieman monimutkaisempaa kuin kiinteällä leveydellä.

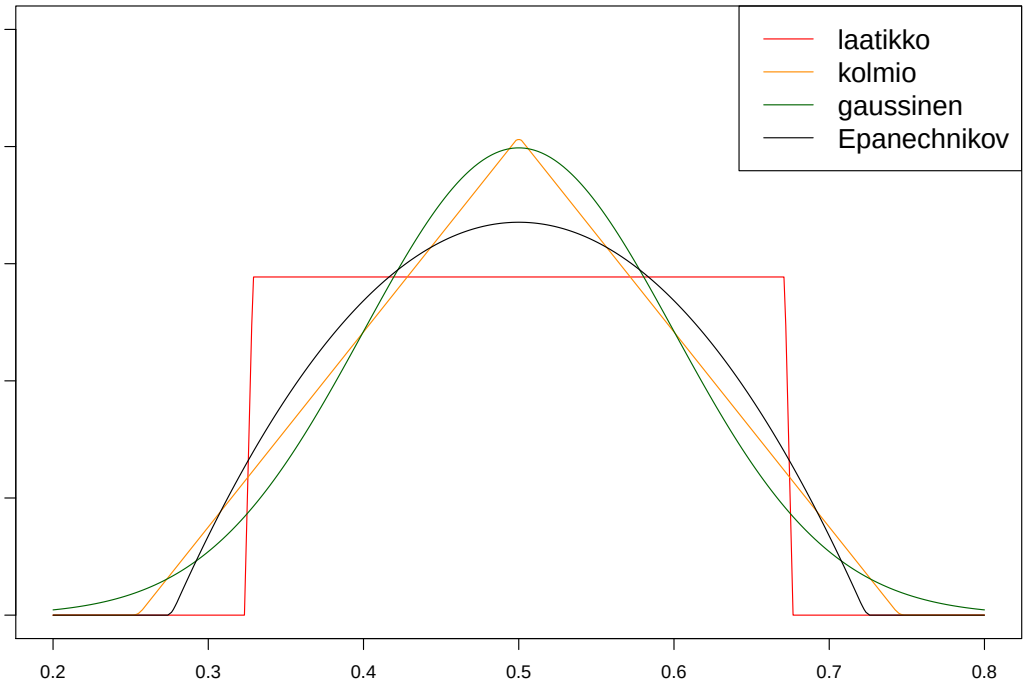
## 3.2 Ydinfunktiot

Ydinestimointiin on olemassa erilaisia ydinfunktioita. Taulukossa 2 on listattuna niistä tavallisimpia ja kuvassa 7 on esitettynä niiden kuvaajat. Kaikki esitetyt ydinfunktiot ovat symmetrisiä nollan suhteen ja kaikkialla ei-negatiivisia.

Taulukko 2: Tavallisimpia ydinestimoinnissa käytettyjä ydinfunktioita.

| *Ydin* | *Ydinfunktio K(x)* |
|---|---|
| Laatikko | $\frac{1}{2}$, jos $\lvert x\rvert < 1$, muulloin $0$ |
| Kolmio | $1 - \lvert x\rvert$, jos $\lvert x\rvert < 1$, muulloin $0$ |
| Gaussinen | $\frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2}$ |
| Epanechnikov | $\frac{3}{4}\left(1 - \frac{1}{5}x^2\right)/\sqrt{5}$, jos $\lvert x\rvert < \sqrt{5}$, muulloin $0$ |

Laatikkoydintä kutsutaan naiiviksi estimaattoriksi. Sitä käyttämällä tiheyskuvaajasta ei tule pehmeä vaan porrasmainen (Silverman, 1986), mutta se on todella yksinkertainen toteuttaa. Epanechnikovin ydintä (Epanechnikov, 1969) pidetään optimaalisena




Kuva 7: Kuvassa on esitetty taulukossa 2 esitetyt ydinfunktiot kuvaajina. Ytimen leveys on 0,1. Tässä tapauksessa on yksi havainto asetettu pisteen 0,5 kohdalle, ja sille on laskettu ydinestimaatti eri ydinfunktioilla.

(Silverman, 1986) ja sen nopea nollautuminen pienentää ytimen vaikutusaluetta verrattuna gaussiseen ytimeen. Gaussinen ydin ei teoriassa ikinä täysin nollaudu, joten sen vaikutus ulottuu kauaskin varsinaisesta havainnosta. Käytännössä sen hännät voidaan kuitenkin katkaista helpottamaan laskentaa (ks. esim. Ramachandran ja Perkins, 2013).

On kuitenkin huomattu ettei edes laatikkoydin ole paljon Epanechnikovin ydintä huonompi todellisen tiheysjakauman esittämisessä (Silverman, 1986). Ydinfunktion valinta ei siis ole tuloksen kannalta niin tärkeä tekijä kuin ytimen leveyden valinta. Ytimen valinnassa voi kuitenkin halutessaan miettiä muita ominaisuuksia, kuten esimerkiksi laskennallista vaativuutta (Silverman, 1986).

Tässä esitetyt ydinfunktiot ovat yleisimmin käytettyjä, mutta ydinfunktioita on muitakin ja tarvittaessa ydinfunktioita voi muodostaa myös itse. Ydinfunktioiden ei myöskään aina tarvitse olla positiivisia ja/tai symmetrisiä (Silverman, 1986) ja joskus kyseisistä ehdoista voi olla tarvetta joustaa.

## 3.3 Ytimen leveyden valinta

Ytimen leveyden valintaan on olemassa joitakin yleisiä valintaperiaatteita, joiden lisäksi on kehitetty myös automaattisia leveydenmääritysmenetelmiä.

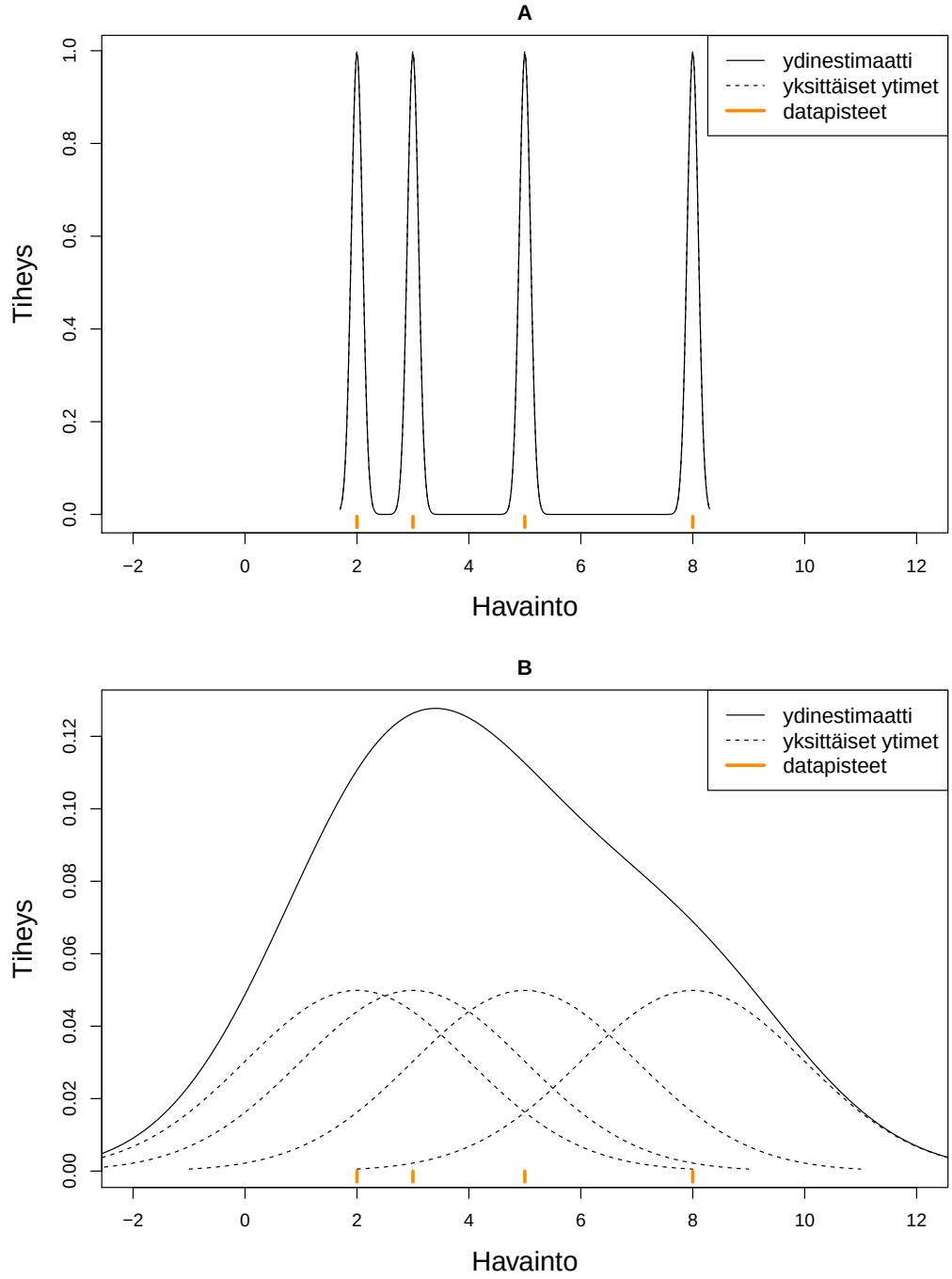
### 3.3.1 Yleiset valintaperiaatteet

Ydinestimoinnin tulokseen vaikuttaa eniten ytimen leveys ja siksi ytimen leveyden valinta on erittäin tärkeä osa ydinestimointia (Sheather, 2004). Liian kapea ydin *alipehmentää* estimaattia (undersmoothing), mikä tekee ydinestimaattorista piikikkään. Liian leveä ydin puolestaan *ylipehmentää* ydinestimaattia (oversmoothing), mikä peittää kaikki datan yksityiskohdat. Kuvassa 8 on esitettynä nämä kaksi eri ääritapausta samoilla datapistellä. Kuvassa 8A on ihan liian kapea ydin, mikä tekee ydinestimaatista piikikkään ja täysin yksittäisten ytimien mukaisen. Kuvassa 8B on puolestaan liian leveä ydin, mikä tekee havainnoista ydinestimaatissa samaa massaa hävittäen yksityiskohtia. Perusteltaessa tuloksia estimaateilla on kuitenkin syytä varoa pehmentämästä liikaa, koska lukija pystyy silmänvaraisesti tasoittamaan enemmän, mutta ei toisinpäin (Silverman, 1986).

Käyttäjä voi itse valita ytimen leveyden, mutta on olemassa myös automaattisia menetelmiä sopivan ytimen leveyden määrittämiseksi. Käyttäjän itse tekemät valinnat voivat olla hyviä tilanteissa, joissa on jo arvioita siitä millainen datan tiheyden muoto on (Silverman, 1986). Tällaisessa tapauksessa tarkastellaan joitakin eri leveyksillä laskettuja tiheyksiä, joiden perusteella valinta tehdään. Automaattiset menetelmät on kehitetty helpottamaan ydinestimoinnin käyttöä, vähentämään tarvetta valita leveys sen mukaan mikä “näyttää hyvältä” sekä parantamaan aloittelijoiden tekemiä estimaatteja (Terrell, 1990). Lisäksi automaattiset menetelmät auttavat, kun tehdään isoja määriä analyyseja monelle eri datalle (Silverman, 1986).

### 3.3.2 Automaattiset leveydenmääritysmenetelmät

Automaattisissa menetelmissä määritetään joko yksi kiinteä leveys koko datalle (fixed bandwidth KDE) tai ytimen leveyttä muutetaan perustuen datan paikalliseen tiheyteen (adaptive/variable KDE). Kiinteän leveyden ydinestimoinnissa joudutaan ytimen leveyttä optimoimaan harvojen alueiden alipehmennyksen ja tiheiden alueiden yli-




Kuva 8: Ytimen leveyden vaikutus ydinestimaattiin. A:ssa on kapea ydin (ytimen leveys = 0.1) ja B:ssä leveä ydin (ytimen leveys = 2). A:ssa yksittäiset ytimet jäävät ydinestimaatin alle. Molemmissa tapauksissa on käytetty gaussista ydintä. Huomattavaa on eroavaisuus Kuvaan 6, jossa on käytetty samoja datapisteitä.

pehmennyksen välillä. Tästä syystä automaattisissa menetelmissä joudutaan käyttämään jonkinlaista virhemittaria leveyden oikeellisuuden/virheellisyyden määrittämiseksi. Mukautuvan leveyden ydinestimaateilla vältetään optimointi tiheiden ja harvo-

jen alueiden välillä, koska ytimen leveys on aina paikallinen riippuen datapisteiden tiheydestä.

Menetelmiä automaattiseen leveydenmääritykseen on kehitetty vuosien saatossa useita, mutta optimia ratkaisua määritykseen ei ole vielä kehitetty. Aiheesta löytyy lisätietoa esimerkiksi seuraavista lähteistä Silverman (1986); Jones ym. (1996); Rudzkis ja Kavaliauskas (1998); Sheather (2004); Raykar ja Duraiswami (2006).

Automaattiset leveydenmääritysmenetelmät on yleensä johdettu jostakin virhefunktiosta (Jones ym., 1996). Yleisimmin käytetyt virhefunktiot ovat *integroitu neliövirhe* (ISE, integrated squared error), *integroitu keskineliövirhe* (MISE, mean integrated squared error) ja MISE:stä johdettu *asymptoottinen integroitu keskineliövirhe* (AMISE, asymptotic mean integrated squared error) (Sheather, 2004). Koska ydinestimoitavan datan todellista tiheyttä ei tiedetä vielä leveyttä määritettäessä, joudutaan virhefunktioissa tiheyden paikalla käyttämään approksimaatioita.

Yksinkertaisimmat tavat saada leveys määritettyä ovat niin kutsutut peukalosäännöt, joissa käytetään apuna joitakin datasta saatavia tunnuslukuja leveyden määrittämiseksi. Eräs esimerkki on Silvermanin peukalosääntö

$$h_{Silverman} = 0.9An^{-1/5},$$

jossa A:n tilalle sijoitetaan datan lähekkäisyyttä kuvaavia tunnuslukuja, kuten keskihajonta (Sheather, 2004; Silverman, 1986). Peukalosäännöt kuitenkin tuottavat useimmissa tapauksissa ylipehmennettyjä estimaatteja (Sheather, 2004).

Muita hieman vanhempia tapoja leveyden määrittämiseksi ovat erilaiset ristiinvalidointimenetelmät. Esimerkiksi *pienimmän neliösumman ristiinvalidointi* (least squares cross-validation) perustuu ISE:n kahden ensimmäisen termin minimoimiseen. ISE:n toinen termi vaatii estimaatin syötteenä, joten sen tilalla käytetään arvioita estimaatista jättämällä yksi datapiste kerrallaan pois (Jones ym., 1996; Sheather, 2004). Ristiinvalidoinnilla saa yleensä tulokseksi useita minimikohtia (Sheather, 2004). Useista minikohdista suurimman lokaalin minimin on huomattu toimivan paremmin kuin globaalin minimin (Jones ym., 1996; Rudzkis ja Kavaliauskas, 1998). Käytännössä ristiinvalidoinnit ovat hitaita laskea suurille datamäärille, koska joudutaan laskemaan useita arvioestimaatteja (Sheather, 2004).

Ristiinvalidointien hitauden vuoksi on kehitetty nopeampia menetelmiä. Esimerkki-

nä tällaisista ovat *sijoitusmenetelmät* (plug-in methods), joissa on virhefunktion tuntemattomat muuttujat korvataan estimaateilla. Esimerkiksi AMISE:ssa tuntemattomana muuttujan on tiheysfunktion toinen derivaatta, jolle annetaan arvoksi jokin *pilottiestimaatti* eli helpolla menetelmällä määritetty lähtöarvo. Eri sijoitusmenetelmät eroavat siinä, miten pilottiestimaatin leveys valitaan. Yksi tapa on laskea pilottiestimaatin leveys käyttäen peukalosääntöjä. Sijoitusmenetelmät tuottavat yleensä pehmeämpiä estimaatteja kuin ristiinvalidointimenetelmät (Sheather, 2004)

Automaattisia leveydenmääritysmenetelmiä on paljon enemmän kuin mitä tässä on mainittu. Huomioitavaa kuitenkin on, että leveyden määrittelemiseksi ei ole vuosikymmenien kehittämisen ja keskustelun jälkeenkään löydetty kaikkeen sopivaa ja parasta menetelmää. Yleensä kuitenkin suositellaan laskemaan monta eri estimaattia eri tavoilla ja tarkastelemaan sitten ovatko ne dataan ja käyttötarkoitukseen sopivia (Sheather, 2004).

## 3.4 Ydinestimoinnin käyttö ChIP-seq analyyseissa

Biologisen datan monimutkaisuuden ja määrän vuoksi bioinformatiikassa on siirrytty käyttämään edistyneitä datan analysointi- ja koneoppimismenetelmiä (Tarca ym., 2007). Tämä on johtanut myös ydinestimointia hyväksikäyttäviin bioinformatiikan menetelmiin. ChIP-seq-data-analyysiinkin on kehitetty muutamia ydinestimointia hyväksikäyttäviä työkaluja, jotka ovat keskittyneet piikkien hakuun sekvenssilukemadatasta (Valouev ym., 2008; Boyle ym., 2008; Ramachandran ja Perkins, 2013). Menetelmien käytöstä ja yleisyydestä ei kuitenkaan ole tietoa. Piikkien koostamiseen ei ole aiemmin kehitetty ydinestimointia hyödyntäviä menetelmiä, minkä vuoksi tässä keskitytään melkein samankaltaisiin piikkien hakuun kehitettyihin menetelmiin.

Suurin ongelma ydinestimoinnin käytössä genomiselle datalle on leveyden valinnassa. Kehitetyt automaattiset menetelmät leveyden määrittämiseen eivät toimi, koska data on liian harvaa koko genomin mittakaavassa (Boyle ym., 2008). Tästä syystä on piikkien haussa ehdotettu ytimen leveydeksi kiinteitä arvoja kuten 30 (Valouev ym., 2008) tai etsittävien alueiden suuruudesta laskettuja arvoja (Boyle ym., 2008). Pienehköt kiinteät ytimet tuottavat kuitenkin ongelmia vähäisten lukemien alueilla, joissa myös yksittäiset lukemat saavat enemmän painoarvoa. Ratkaisuksi Ramachandran ja Perkins (2013) ovat kehittäneet mukautuvan ytimen ydinestimointia käyttävän menetelmän, joka ot-

taa huomioon seitsemän lähimmän sekvenssilukeman etäisyyden laskiessaan ytimen leveyttä jokaiselle datapisteelle. Mukautuvan ytimen ydinestimoinnin avulla saadaan siis aina datasta riippuva leveys, joka ei ole liian leveä eikä siten hävitä liikaa yksityiskohtia. Tämän perusteella mukautuvan ytimen menetelmä vaikuttaisi paremmin soveltuvalta sekvenssilukemadatan käsittelyyn.

Pienempi ongelma ydinestimoinnin käyttämiselle on se, että ydinestimointi tarvitsee pistemäistä dataa, jota sekvenssilukemadata ei ole. Tähän ratkaisuna on ollut joko ottaa fragmentin oletettu keskikohta (Boyle ym., 2008) tai sekvenssilukeman alkukohta (Ramachandran ja Perkins, 2013; Valouev ym., 2008) edustamaan kutakin sekvenssilukemaa.

Genomisen datan kanssa haasteena ovat siis samat ongelmat kuin ylipäänsä ydinestimoinnissa. Ydinestimoinnista on kuitenkin niin vähän kokemusta bioinformatiikassa, ettei ongelmiin ole tarjolla mitään systemaattiseen tutkimukseen perustuvia ohjenuoria.

# 4 Uusi menetelmä ChIP-seq-piikkidatan koostamiseen

Tässä luvussa käydään läpi ChIP-seq-piikkien koostamista ja esitellään uusi ydinestimointiin perustuva menetelmä ChIP-seq-piikkien koostamiseen. Lopuksi käydään läpi muita ChIP-seq-piikkien koostamis- ja vertailumenetelmiä.

## 4.1 Koostamisen ja vertailun ero

Tarve ChIP-seq-piikkien koostamis- ja vertailumenetelmille on syntynyt vasta viime aikoina julkisen datan määrän kasvettua. Yleinen käyttötarkoitus julkiselle datalle on niiden piikkien samankaltaisuuden tai päällekkäisyyden tarkastelu vertailemalla piikkejä eri tutkimusten kesken. Iso osa olemassa olevista menetelmistä ja kirjallisuudesta keskittyy näytteiden vertailuun. Tästä syystä on tarpeellista täsmentää eroavaisuuksia vertailun ja koostamisen välillä.

ChIP-seq-piikkien koostamisella tarkoitetaan monesta eri ChIP-seq-tutkimuksesta saatujen piikkien yhdistämistä yhdeksi koosteeksi. Tämä tarkoittaa, että eri tutkimuksista tulleet samalla alueella genomissa sijaitsevat piikit ovat koosteessa yhtenä koostepiikkinä. Koostamisessa pyritään siis saamaan aikaiseksi kaikkia koostamisessa mukana olevia näytteitä kuvaava kooste. Vertailtaessa puolestaan yritetään etsiä tilastollisesti merkitseviä eroavaisuuksia tai samankaltaisuuksia ChIP-seq-piikkien ja tulosten välillä. Tärkeää on huomata, että kooste voi toimia myös pohjana vertailulle ja muille tutkimuksille. Erilaisia koostamis- ja vertailumenetelmiä käsitellään tarkemmin luvussa 4.3.

## 4.2 ConsensusSummit-menetelmä

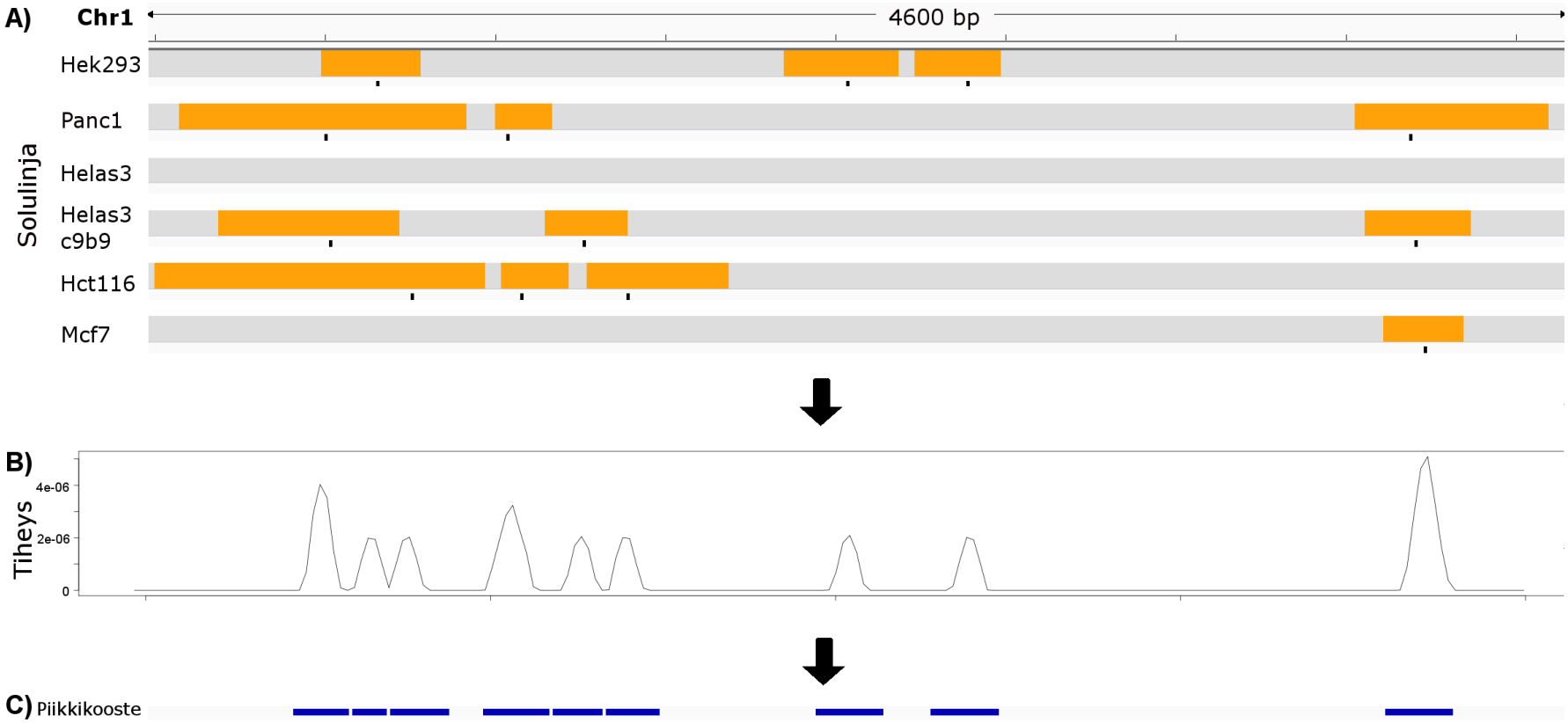
Tämän tutkimuksen tuloksena kehitettiin uusi, ydinestimointiin perustuva ConsensusSummit-menetelmä, joka voi koostaa usean ChIP-seq-kokeen piikkejä. Menetelmää on käytetty aiemmin julkaisussa (Tuoresmäki ym., 2014) D-vitamiinireseptorin sitoutumisen tutkimiseen yhdistämällä usean eri tutkimuksen dataa. Julkaisussa ei kuitenkaan paneuduttu menetelmän yksityiskohtiin kovinkaan tarkasti ja menetelmää on sittemmin kehitetty lisää.

ConsensusSummit-menetelmä koostuu viidestä vaiheesta: 1) Esiprosessoinnista 2) Ydinestimoinnista, 3) Ydinestimaatin maksimikohtien etsinnästä, 4) Koostepiikkien muodostuksesta sekä 5) Loppuprosessoinnista. ConsensusSummit-menetelmän periaate on esitetty Kuvassa 9. Perusidea on seuraava: Piikkien (Kuva 9A, oranssit laatikot) huippupisteistä (Kuva 9A, mustat pisteet) lasketaan ydinestimaatti (Kuva 9B), jonka maksimikohdista muodostetaan koostepiikit piikkikoosteeseen (Kuva 9C). Välituloksena saatava ydinestimaatti on kooste jo itsessään, mutta hakemalla maksimikohdat saadan koostepiikin keskustalle tarkempi sijainti.

ConsensusSummit-menetelmän syötteenä ovat ChIP-seq tutkimuksista analysoidut piikit sekä ytimen leveys, käytettävä ydinfunktio, diskretisoinnin ikkunakoko sekä tutkittavan eliön kromosomien pituudet. Ensimmäisessä vaiheessa eli esiprosessoinnissa data muunnetaan ConsensusSummit-menetelmän vaatimaan muotoon eli ChIP-seq-piikeistä kerätään tiedot piikkien kromosomeista ja huippupisteiden sijainneista. Menetelmässä hyödynnetään vain huippupistettä, koska ydinestimointi vaatii pistemäistä dataa. Huippupiste on tähän tarkoitukseen hyvä, koska se on valmiiksi pistemäinen sijainti ja se vastaa suurimman signaalin omaavaa nukleotidia kussakin piikissä. Esimerkiksi kahden huippupistedatan (Taulukot 3 ja 4) yhdistelmänä saatu data on esitetty Taulukossa 5.

Toisessa vaiheessa käytetään ydinestimointia luomaan estimaatti piikkien huippupisteiden tiheydestä (Algoritmi 1). Piikkien huippupisteiden tiheyden avulla voidaan määritellä koostepiikkien sijainnit kromosomeittain. Ydinestimointi tehdään kromosomeittain, koska kromosomit ovat fyysisestikin erillisiä ja niillä on jokaisella oma nukleotidikoordinaatistonsa. Esimerkiksi merkintä chr1:553021 tarkoittaa kromosomin 1 nukleotidiä 553021 ja merkintä chr2:553021 vastaavasti kromosomin 2 nukleotidiä 553021. Kromosomit ovat myös erittäin pitkiä (esimerkiksi ihmisellä kymmeniä miljoonia nukleotidejä), minkä takia ydinestimaatin laskeminen ja käsittely jokaiselle nukleotidille on hankalaa. Suurta kokoa pienennetään diskretisoimalla kromosomit tietyn nukleotidimäärän mittaisiksi ikkunoiksi. Käyttäjä määrää diskretisoinnin määrän. Jokaisen huippupisteen sijainti sekä ydinestimoinnin ytimen leveys käsitellään diskretisoinnin muodostamien ikkunoiden tarkkuudella.

Kolmannessa vaiheessa diskretisoidusta ydinestimaatista etsitään maksimikohdat. Maksimikohtien etsintään käytetään Algoritmia 2. Maksimikohdat etsitään ns. liukavalla ikkunalla diskretisoidusta ydinestimaatista. Liukavan ikkunan leveyden käyttäjä saa valita. Jos käyttäjä on esimerkiksi valinnut ydinestimaatin diskretisoimisen 20




Kuva 9: Esimerkki ChIP-seq-piikkien koostamisesta ConsensusSummit-menetelmällä. ChIP-seq-piikeistä saatujen huippupisteiden (A, mustalla) ydinestimaatin (B) paikallisten maksimien eli estimaatin huippupisteiden avulla muodostetaan koostepiikit (C). Kuvassa A, B ja C ovat keskenään linjassa.

Taulukko 3: Kuvitteellinen esimerkki piikeistä ChIP-seq-kokeessa 1.

| *Kromosomi* | *Alku* | *Loppu* | *Pituus* | *Huippupiste* | *Nimi* |
|---|---|---|---|---|---|
| chr1 | 852038 | 852337 | 300 | 852113 | Hek293_piikki1 |
| chr1 | 877029 | 877470 | 442 | 877315 | Hek293_piikki2 |
| chr11 | 114835604 | 114835890 | 287 | 114835748 | Hek293_piikki3 |
| chr20 | 307885 | 308140 | 256 | 307983 | Hek293_piikki3 |
| chrX | 10584575 | 10584745 | 171 | 10584642 | Hek293_piikki4 |

Taulukko 4: Kuvitteellinen esimerkki piikeistä ChIP-seq-kokeessa 2.

| *Kromosomi* | *Alku* | *Loppu* | *Pituus* | *Huippupiste* | *Nimi* |
|---|---|---|---|---|---|
| chr1 | 877231 | 877401 | 171 | 877314 | Mcf7_piikki1 |
| chr10 | 21907990 | 21908156 | 167 | 21908090 | Mcf7_piikki2 |
| chr12 | 3002257 | 3002475 | 219 | 3002359 | Mcf7_piikki3 |
| chr20 | 599628 | 599949 | 322 | 599791 | Mcf7_piikki4 |
| chr21 | 15269032 | 15269210 | 179 | 15269120 | Mcf7_piikki5 |

Taulukko 5: Esimerkki kokeista 1 ja 2 kerätyistä huippupisteistä ydinestimointia varten.

| *Kromosomi* | *Huippupisteet* |
|---|---|
| chr1 | 852113, 877314, 877315 |
| chr10 | 21908090 |
| chr11 | 114835748 |
| chr12 | 3002359 |
| chr20 | 307983, 599791 |
| chr21 | 15269120 |
| chrX | 10584642 |

**Algoritmi 1** Ydinestimaatin muodostus

**Syöte:**

$w$ *// Ikkunan leveys*

$h$ *// Ytimen leveys*

$K$ *// Ydinfunktio*

$cl = \{l_1, ..., l_N\}$ *// Kromosomien pituudet*

*// X:t ovat kromosomin huippupisteet*

$huiput[1] = \{X_{1,1}, ..., X_{1,lkm_1}\}$

$\vdots$

$huiput[N] = \{X_{N,1}, ..., X_{N,lkm_N}\}$

**Tuloste:**

*// wlkm = kyseisen kromosomin pituus / w*

*// ja e on estimaatin arvo kyseisessä ikkunassa*

$e[1] = \{e_{1,1}, ..., e_{1,wlkm_1}\}$

$\vdots$

$e[N] = \{e_{N,1}, ..., e_{N,wlkm_N}\}$

**Metodi:**

**for** $i = 1$ **to** $N$ **do**

  $y = len(huiput[i]$ ) *// Huippupisteiden lukumäärä kromosomissa i*

  **for** $k = 1$ **to** $y$ **do**

    *//ikkos() muuntaa huipun sijainnin vastaavaksi ikkunaksi*

    $huiput[i][k] = ikkos(huiput[i][k], w)$

  **end for**

  **for** $j = 1$ **to** $cl[i] \div w$ **do**

    $e[i][j] = \frac{1}{y \times h} \sum_{k=1}^{y} K\left(\frac{j - huiput[k]}{h}\right)$

  **end for**

**end for**

**return** $e$

nukleotidin tarkkuuteen ja maksimin etsinnässä käytettävän liukuvan ikkunan kooksi 10, etsitään maksimia nukleotideissä mitattuna silloin 200 nukleotidin alueelta. Koska maksimin etsinnässä käytetään diskretisoitua estimaattia, on sen antama maksimikohtakin nukleotideissä mitattuna vain diskretisoinnin tarkkuudella oikein. Liukuvasta ikkunasta johtuen maksimikohtien ja siten myös koostehuippujen etäisyys toisistaan voi olla minimissään puolet maksimi-ikkunan koosta.

**Algoritmi 2** Maksimikohtien etsintä ydinestimaatista

**Syöte:**
$e$ *// Algoritmin 1 tuloste*
$m$ *// Liukuvan ikkunan leveys*

**Tuloste:**
*// M:t ovat estimaatin huippukohtia kyseisessä kromosomissa*
$huippukohdat[1] = \{M_{1.1}, ..., M_{1.lkm_1}\}$
$\vdots$
$huippukohdat[N] = \{M_{N.1}, ..., M_{N.lkm_N}\}$

**Metodi:**
**for** $i = 1$ **to** $N$ **do**
  **for** $alku = 1$ **to** $len(e[i]) - m$ **do**
    $loppu = alku + m$
    $tmp = maksimikohta(e[i], alku, loppu)$
    **if** $((tmp > 0)$ **and** $(tmp \neq loppu)$ **and** $(tmp \notin huippukohdat[i]))$ **then**
      *lisää tmp huippukohdat[i]:hin*
    **end if**
  **end for**
**end for**
**return** $huippukohdat$

**Apufunktio:**
$maksimikohta(taulukko, alku, loppu)$
*// Palauttaa 0, jos maksimi ≤ 0 koko taulukossa*
$y = 0$
$max = 0$
**for** $i = alku$ **to** $loppu$ **do**
  **if** $taulukko[i] \geq max$ **then**
    $y = i$
  **end if**
**end for**
**return** y

Koska ydinestimaatti kuvaa huippupisteiden tiheyttä, ovat estimaatin maksimikohdat

lähellä piikkien huippupisteitä. Näin ollen mitä suurempi estimaatin maksimiarvo on, sitä enemmän estimaatti-ikkunan lähellä on piikkien huippupisteitä. Saaduista maksimikohdista tehdään koostepiikkien keskikohtia. Koostepiikkien keskikohdista muodostetaan piikkejä lisäämällä niihin käyttäjän valitsema määrä nukleotidejä molempiin suuntiin. Jos käyttäjällä on esimerkiksi koostepiikin keskus chr1:55510 ja käytetään sadan nukleotidin levennystä, tulee kyseisen koostepiikin sijainniksi chr1:55410-55610. Koostepiikin leveys voi riippua myös valitusta ytimen leveydestä.

Loppuprosessointivaiheessa poistetaan koostepiikkien päällekkäisyyksiä ja lisätään koostepiikkeihin tietoa. Päällekkäisyys voi tuottaa hankaluuksia myöhemmin, joten piikkien päällekkäisyys täytyy poistaa. Päällekkäisyys poistetaan katkaisemalla päällekkäiset koostepiikit päällekkäisen alueen puolivälistä ja jakamalla katkaistu osa koostepiikkien kesken. Esimerkiksi jos kyseessä olisivat päällekkäiset piikit chr1: 55410-55610 ja chr1: 55600-55800, tulee niistä päällekkäisyyden poiston jälkeen chr1: 55410-55605 ja chr1: 55606-55800. Koostepiikin keskustan sijaintiin ei puututa. Tiedon lisäyksellä puolestaan tarkoitetaan alkuperäisten piikkitietojen lisäämistä koostepiikkeihin. Tämä tapahtuu esimerkiksi etsimällä koostepiikin alueelle osuvat alkuperäisten piikkien huippupisteet. Näin saadaan selville mikä huippupiste vaikutti mihinkin koostepiikkiin. Alkuperäisten huippupisteiden avulla voidaan puolestaan kerätä muitakin tietoja alkuperäisistä piikeistä, esimerkiksi piikin leveys tai lukemasignaalin voimakkuus. Tarvittavat tiedot voi valita sen mukaan, mitä koosteella on tarkoitus tehdä.

ConsensusSummit-menetelmän hyvänä puolena on, että kaikki koosteeseen käytettävät ChIP-seq-kokeet voidaan käsitellä samanaikaisesti. Lisäksi ChIP-seq-kokeiden määrä ei vaikuta laskenta-aikaan kovinkaan merkittävästi, koska kokeet lisäävät vain piikkien huippupisteitä ydinestimaattiin, joka joudutaan huippupisteiden määrästä riippumatta laskemaan koko genomille. ConsensusSummit-menetelmässä on myös puutteita. Eräs rajoitus on diskretisoinnista johtuva epätarkkuus, jota voi yrittää vähentää pienemmällä diskretisointivälillä, mutta tämä kuluttaa enemmän muistia. Toinen vaihtoehto epätarkkuuden pienentämiseksi on laskea koostepiikkiin kuuluvien alkuperäisten huippupisteiden sijaintien keskiarvo ja käyttää sitä maksimikohdan sijasta. Toinen menetelmän rajoitus on ettei se salli tulosten päivittämistä jälkikäteen uusilla ChIP-seq-piikkidatoilla. Jos uusien ChIP-seq-kokeiden tuloksia halutaan lisätä koosteeseen, täytyy koosteen laskenta suorittaa kokonaan uudestaan.

## 4.3 Muita koostamis- ja vertailumenetelmiä

Varsinaisia koostamismenetelmiä on olemassa todella vähän, mutta kahden eri ChIP-seq-kokeen tulosten vertailumenetelmiä on useampia. Koostamista on tiettävästi tehnyt vain ENCODE-projektin konsortio (ENCODE Project Consortium, 2012). Konsortio on tehnyt jokaiselle projektissa tutkitulle transkriptiofaktorille koosteen kyseisen transkriptiofaktorin sitoutumispaikoista genomissa. Sitoutumispaikat on kuitenkin esitetty koosteessa karkealla resoluutiolla eli ne kertovat transkriptiofaktorin sitoutumisalueen vain suuntaa-antavasti. Koosteet ovat julkisesti saatavilla ja niitä käytetään myöhemmin vertailuun tässä esitettävän ConsensusSummit-menetelmän kanssa (Luku 5).

ChIP-seq-kokeiden välisiä vertailumenetelmiä on sen sijaan monia. Eri vertailumenetelmillä pyritään esimerkiksi määrittämään transkriptiofaktorin sitoutumisen samankaltaisuutta eri soluissa tai selvittämään kahden eri transkriptiofaktorin sitoutumisen riippuvutta toisistaan. Kaikki vertailumenetelmät perustuvat parittaisiin vertailuihin ChIP-seq-kokeiden kesken (ks. esim. Maehara ym., 2012; Shao ym., 2012; Chikina ja Troyanskaya, 2012). Vertailut ovat osoittautuneet haasteelliseksi sekä ChIP-seq-menetelmästä (Chen ym., 2015) että analyysitavoista johtuvista syistä (Shao ym., 2012). Analyysiperäisistä syistä keskeisin on itse piikin määritelmä, koska piikkejä tunnistaessa joudutaan valitsemaan piikin $p$-arvolle jokin raja. Raja-arvosta johtuen voi toisesta näytteestä tulla piikki ja toisesta ei, vaikka näytteiden välillä ei olisi suurta eroa (Shao ym., 2012). Tämä ongelma on vaikein kun halutaan vertailla useita näytteitä keskenään.

Perinteinen menetelmä kahden ChIP-seq-kokeen vertailuun on laskea piikkien päällekkäisten nukleotidien määrä joko absoluuttisesti tai suhteutettuna piikin leveyteen (ks. esim. Johnson ym., 2007). Tähän vertailumenetelmään vaikuttaa esimerkiksi piikkien leveys, joka voi vaihdella paljonkin piikkien hakumenetelmästä riippuen (Chikina ja Troyanskaya, 2012). Perinteisen tavan heikkouksien vuoksi on kehitetty parempia menetelmiä, jotka pureutuvat erilaisten ongelmien ratkaisuun.

ChIP-seq-kokeiden samankaltaisuuden mittaamiseen sopii esimerkiksi Chikinan ja Troyanskayan (2012) kehittämä menetelmä, joka perustuu etäisyyksien ja päällekkäisyyden tarkasteluun verrattavien ChIP-seq-piikkidatojen piikkien kesken. Menetelmässä pyritään vähentämään piikkien leveyden vaikutusta ottamalla aina lähimmät piikit tarkasteluun riippumatta siitä ovatko ne päällekkäin vai eivät. Menetelmän tuloksena saadaan $p$-arvojakauma piikkien lähekkäisyydestä. Jakauman perusteella voi tehdä

johtopäätöksiä näytteiden samankaltaisuudesta.

Toinen samankaltaisuutta mittaava menetelmä on esimerkiksi MAnorm (Shao ym., 2012). MAnorm perustuu ChIP-seq-piikkien sekvenssilukemasignaalin normalisointiin kahden ChIP-seq-kokeen kesken. MAnorm käyttää normalisoinnin apuna verrattavien datojen yhteisiä eli samassa sijainnissa olevia ChIP-seq-piikkejä. Normalisointi tehdään skaalaamalla molempien kokeiden lukemasignaalit yhteisten piikkien lukemasignaalien voimakkuuserojen ja ChIP-seq-kokeiden sekvenssilukemien kokonaismäärien suhteiden avulla. Normalisoitujen lukemasignaalien vuoksi MAnorm:lla voidaan tutkia myös yksittäisten ChIP-seq-piikkien välisiä sitoutumiseroja.

Vertailumenetelmillä voidaan tutkia myös eri transkriptiofaktoreiden välisiä riippuvuussuhteita. Esimerkiksi Maehara ym. (2012) kehittämä menetelmä käyttää riippuvuuden tutkimiseen kahden ChIP-seq-kokeen lähimmäisten piikkien välistä etäisyyshajontaa. Etäisyyshajontaan lasketaan etäisyys jokaisesta verrattavan ChIP-seq-kokeen piikistä lähimpään piikkiin verrokkikokeessa. Piikkien etäisyyshajonnan muodon perusteella voidaan päätellä esimerkiksi onko verrattavan transkriptiofaktorin sitoutuminen verrokista riippumatonta, verrokin sitoutumista tehostavaa vai verrokin sitoutumista heikentävää.

Vertailumenetelmiä ei kuitenkaan ole yleensä suunniteltu useamman kuin kahden ChIP-seq-kokeen tapauksia varten. Parittaisten vertailujen tekeminen usean eri kokeen kesken on työlästä, mitä voisi helpottaa tekemällä yhteisen koosteen vertailun pohjaksi. ChIPComp (Chen ym., 2015) on menetelmä, joka tekee ensin yhdisteen (unioni) kaikista vertailuun käytettävien ChIP-seq-kokeiden piikkialueista. Yhdisteen muodostamistapaa ei ole artikkelissa tarkemmin kuvattu. Yhdistettä verrataan pareittain kaikkiin tutkimuksen ChIP-seq-kokeiden piikkeihin, mikä vähentää tehtävien vertailujen määrää huomattavasti. ChIPComp:in yhdistettä voi pitää eräänlaisena koosteena. Koosteen avulla saadaan vertailuun mukaan kaikki genomiset sijainnit, joissa edes yhdessä näytteessä on piikki. Pelkästään kahden näytteen vertailuun kehitetyt menetelmät ohittavat tämän yhteisen koosteen tekemisen.

ChIPComp eroaa ConsensusSummit-menetelmästä siinä, että se käyttää unionin tekemiseen koko piikkiä, piikkien huippupisteiden sijasta. ChIPComp:in käyttötarkoitus on myös suppeampi kuin ConsensusSummit-menetelmän, koska ChIPComp:in unioni toimii vain menetelmän omana aputyökaluna. ConsensusSummit-menetelmän tavoitteena on kuitenkin olla hyödyllinen myös esimerkiksi motiivien sitoutumisen tar-

kastelussa, mikä vaatii koosteelta tarkempaa sitoutumispaikkatietoa kuin kokonaisten piikkien unioni. Koska piikkien huippupisteitä pidetään transkriptiofaktorin varsinaisina sitoutumispaikkoina, saadaan huippupisteiden avulla tehdystä koosteesta tarkkaa tietoa varsinaisesta sitoutumispaikasta.

# 5 ConsensusSummit-menetelmän empiirinen testaus

ConsensusSummit-menetelmää ei ole analysoitu laajemmin, joten on syytä tarkastella sen toimivuutta ja hyötyjä. Tässä luvussa on tarkoituksena määrittää sopivia parametreja ConsensusSummit-menetelmälle sekä testata menetelmän toimintaa esimerkiksi vertaamalla menetelmän tuottamia tuloksia vertailukelpoiseen aineistoon.

## 5.1 Tavoitteet ja menetelmät

Testauksessa oli kaksi eri kokonaisuutta. Ensimmäinen kokonaisuus keskittyi menetelmän parametreihin ja toinen kokonaisuus tulosten analysointiin sekä vertailuun muiden tutkimusten kanssa.

Parametrien testauksessa tavoitteena oli selvittää hyviä ohjenuoria menetelmän parametrien valintaan sekä tarkastella parametrien vaikutusta tuloksiin. Parametreissa tutkitaan erityisesti ydinestimoinnin ytimen leveyttä, eri ydinfunktioita sekä tuloksena saatavan koostepiikin leveyttä. Koostepiikin leveys on vaikea määrittää tarkasti, joten ensin tutkittiin miten monikertainen koostepiikin leveys kannattaa valita ytimen leveyteen nähden. Testissä koostepiikkien leveys pidettiin 200 emäsparissa ja gaussisen ytimen leveyttä kasvatettiin 20:stä emäsparista 50:een emäspariin. Koostepiikkien leveyden valinnan tulosta käytettiin hyödyksi, kun testattiin ytimen ja ytimen leveyden vaikutusta tuloksiin. Ytimen leveys testattiin erikseen gaussisella, Epanechnikov- sekä kolmioytimellä. Ytimen leveyttä kasvatettiin 20:stä 150:een emäspariin ja koostepiikkien leveys pidettiin kymmenkertaisena ytimen leveyteen nähden.

Mittareina parametrien testauksessa käytettiin koostepiikkien, päällekkäisten koostepiikkien sekä koostepiikkien alkuperäisten huippupisteiden lukumääriä. Koostepiikkien lukumäärässä pyrittiin mahdollisimman pieneen lukuun pitäen huippupisteiden määrän eri piikeissä mahdollisimman korkeana. Samalla päällekkäisiä koostepiikkejä pitäisi olla mahdollisimman vähän ja piikittömiä huippupisteitä ei mielellään ollenkaan. Näillä mittareilla pyrittiin saamaan koostaminen yhdistämään piikkejä mahdollisimman hyvin, kuitenkaan tekemättä koostepiikeistä leveämpiä tai kapeampia kuin on tarpeen.

Toisena kokonaisuutena oli menetelmän tuottamien tuloksien tarkastelu, joka oli jaettu kahteen osaan. Ensimmäisenä keskityttiin sitoutumismotiiveihin. Koska testausdata

on TCF7L2-transkriptiofaktorin (vanhalta nimeltään TCF4) ChIP-seq-dataa, tutkittiin TCF7L2:n motiivin esiintyvyyttä. Motiivi on esitetty aiemmin Kuvassa 5. Sitoutumismotiiveja pidetään transkriptiofaktorin varsinaisena sitoutumispaikkana, joten motiivin esiintymistä piikeissä käytetään usein perustelemaan piikin oikeellisuutta. Tästä syystä tarkasteltiin motiivien esiintyvyyttä koostepiikeissä ja verrattiin miten lähellä motiivit olivat koostepiikkien keskustaa verrattuna alkuperäisten piikkien huippukohtiin. Etäisyydellä koostepiikin keskustaan pyrittiin selvittämään tarkentaako koostaminen piikkejä kohti motiiveja. Samalla tarkasteltiin yleisesti sitoutumisen riippuvuutta motiivista. TCF7L2 motiivin etsintään piikeistä käytettiin HOMER-työkalua (Heinz ym., 2010), jonka mukana tulleistä motiivitiedostoista käytettiin TCF4:n motiivia.

Toisen kokonaisuuden toisessa osassa verrattiin tulosten yhdenmukaisuutta ENCODE:n (ENCODE Project Consortium, 2012) tuottamien sitoutumisalueiden kanssa. Vertailussa keskityttiin erityisesti ENCODE:n sitoutumisalueiden ja tässä muodostettujen koostepiikkien päällekkäisyyteen. Vertailuun on hyvät edellytykset, koska molemmat vertailtavat alueet/piikit on muodostettu samasta datasta. Päällekkäisyyttä tutkittiin sekä nukleotidien että sitoutumisalueiden/koostepiikkien tarkkuudella. Yhdenmukaisuuden mittarina käytettiin kaavoja

$$\alpha = \frac{W\left(P-, E+\right)}{W\left(E+\right)} \quad \text{ja} \tag{1}$$

$$\beta = \frac{W\left(P+, E-\right)}{W\left(P+\right)}, \tag{2}$$

jossa $W\left(P-, E+\right)$ on alueiden/nukleotidien määrä, joissa on ENCODE-alue ilman ConsensusSummit-koostepiikkiä, $W\left(E+\right)$ on ENCODE-alueiden/nukleotidien kokonaismäärä, $W\left(P+, E-\right)$ on alueiden/nukleotidien määrä, joissa on ConsensusSummit-koostepiikki ilman ENCODE-aluetta ja $W\left(P+\right)$ on koostepiikkien tai niiden nukleotidien kokonaismäärä. Lisäksi tutkittiin sellaisia alueita, joissa on vain joko ENCODE:n alue tai tässä tutkimuksessa tuotettu koostepiikki.

Testausta varten ConsensusSummit-menetelmä toteutettiin R-ohjelmointikielellä (R Core Team, 2014), jonka valmiit funktiot nopeuttivat menetelmän toteutusta.

## 5.2 Datan kuvaus

Testaukseen käytetty ChIP-seq-raakadata koostui ENCODE-projektin TCF7L2-transkriptiofaktorin seitsemästä eri solulinjasta. Datat ovat saatavilla NCBI:n SRA-tietokannasta (Ks. Liite 1) ja tunnistenumerot datan löytämiseksi on esitetty Taulukossa 6. TCF7L2 valittiin sekä tutkimusryhmän kiinnostuksesta kyseistä transkriptiofaktoria kohtaan että vertailukelpoisen tulosten olemassaolon vuoksi. Raakadatojen solutyypit, sekvenssilukemien määrät ja niitä kuvaavat parametrit on esitetty Taulukoissa 7 (Näytteet) ja 8 (Kontrollit). Parametreinä näytteille ovat sekvenssilukemien rinnastuvuus genomiin, piikkien määrä, piikkien keskietäisyys seuraavaan piikkiin sekä piikkien keskipituus. Kontrollinäytteille on sekvenssilukemien lisäksi ilmoitettu vain kontrollityyppi sekä sekvenssilukemien rinnastuvuus genomiin.

Taulukko 6: Testauksessa käytettyjen ChIP-seq-datojen SRA-tietokannan GSM-tunnistenumerot.

| *Solulinja* | *Näyte-GSM* | *Kontrolli-GSM* |
|---|---|---|
| HepG2 | GSM782122 | GSM017343 |
| HCT-116 | GSM782123 | GSM817344 |
| HEK293 | GSM782124 | GSM935586 |
| Hela-S3 | GSM816436 | GSM818744 |
| Hela-S3 c9b9 | GSM935625 | GSM818744 |
| PANC-1 | GSM816437 | GSM935617 |
| MCF-7 | GSM816438 | GSM935485 |

Raakadata rinnastettiin genomiin käyttäen Bowtie-työkalun (Langmead ym., 2009) versiota 1.0.1. Rinnastusvaiheessa poistettiin kaikki sekvenssilukemat, jotka eivät rinnastuneet genomiin tai rinnastuivat moneen paikkaan genomissa. Taulukoissa 7 ja 8 on ilmoitettu, kuinka monta prosenttia alkuperäisistä lukemista jäi kyseisten poistojen jälkeen jäljelle.

Piikkien hakuun käytettiin MACS-työkalun (Zhang ym., 2008) versiota 2 (MACS2). Piikkien hakua varten sekä kontrolli- että näytereplikaatit yhdistettiin keskenään. Ennen piikkien hakua poistettiin samasta kohdasta alkavat lukemat käyttäen MACS2 oletusarvoja. Sekvenssilukemien määrä myös skaalattiin MACS2-oletuksia käyttäen näytteen ja kontrollin välillä pienempään määrään. MACS2:ssa piikkien hyväksimisrajana käytettiin parametria $q < 0{,}01$. Piikkien määrät ja niistä lasketut lukuarvot on esitetty Taulukossa 7. Piikkejä saatiin yhteensä 77881 kappaletta. Näistä vielä poistettiin mitokondriaalisen DNA:n piikit, joita oli yhteensä 7 kappaletta. Koostamista varten saatiin

Taulukko 7: Näytteiden tietoja. Eri replikaattien arvot on erotettu puolipisteellä. Replikaatit yhdistettiin ennen piikkien hakua. Keskietäisyys on laskettu piikistä seuraavaan piikkiin. Lukemat rinnastettiin bowtie-työkalulla (Langmead ym., 2009) ja piikit on haettu MACS2-työkalulla (Zhang ym., 2008).

| *Solulinja* | *Solutyyppi* | *Replikaatteja* | *Lukemia* | *Rinnastuvuus* | *Piikkejä* | *Keskietäisyys (ep)* | *Keskipituus (ep)* |
|---|---|---|---|---|---|---|---|
| HepG2 | Maksa | 2 | 26710367; 25005577 | 77,22 %; 76,81 % | 1725 | $1646114,1 \pm 2700144,0$ | $330,0 \pm 154,7$ |
| HCT-116 | Suolisto | 2 | 12217110; 6689144 | 71,54 %; 71,22 % | 29279 | $100176,5 \pm 371972,7$ | $318,3 \pm 194,5$ |
| HEK293 | Munuainen | 2 | 30901682; 26928039 | 78,83 %; 77,76 % | 7718 | $378381,6 \pm 810051,6$ | $334,9 \pm 167,4$ |
| HeLa-S3 | Kohdunkaula | 2 | 19726070; 12699914 | 69,55 %; 74,68 % | 2435 | $1189139,2 \pm 2031883,1$ | $339,4 \pm 159,9$ |
| HeLa-S3 c9b9 | Kohdunkaula | 2 | 31758574; 39988209 | 76,00 %; 71,00 % | 11971 | $172385,1 \pm 523178,0$ | $367,1 \pm 182,3$ |
| PANC-1 | Haima | 2 | 17252445; 24170232 | 71,87 %; 71,50 % | 7775 | $244288,5 \pm 659390,2$ | $366,9 \pm 228,6$ |
| MCF-7 | Rinta | 2 | 26627932; 28223626 | 71,25 %; 75,50 % | 16978 | $374231,1 \pm 805768,1$ | $313,1 \pm 126,1$ |

Taulukko 8: Kontrollinäytteiden tiedot. *: Sama kontrolli molemmille HeLa-S3 näytteille.

| *Solulinja* | *Tyyppi* | *Replikaatteja* | *Lukemia* | *Rinnastuvuus* |
|---|---|---|---|---|
| HepG2 | Input | 1 | 28007793 | 57,39 % |
| HCT-116 | Input | 1 | 11126747 | 64,68 % |
| HEK293 | Input | 1 | 30898350 | 64,41 % |
| HeLa-S3* | Input | 1 | 27143615 | 72,74 % |
| PANC-1 | Input | 1 | 41671673 | 60,84 % |
| MCF-7 | Input | 2 | 18714964; 24478545 | 55,30 %; 55,70 % |

siis yhteensä 77874 piikkiä.

## 5.3 Tulokset

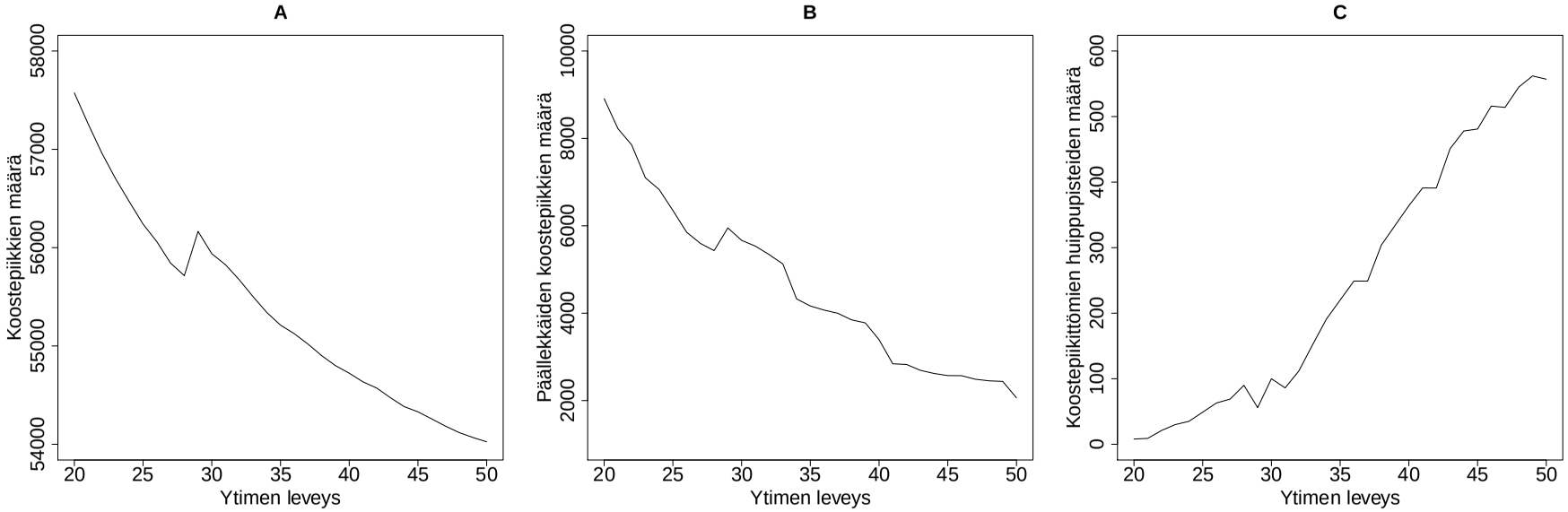
Tuloksissa käydään ensin läpi parametrien valinta ja niiden vaikutus, jonka jälkeen esitetään motiivihaun ja ENCODE yhteenmukaisuuden tulokset.

### 5.3.1 Parametrien vaikutus tuloksiin

Ensimmäisenä testinä oli määrittää miten monikertainen koostepiikin leveyden kannattaa olla ytimen leveyteen nähden. Tulokset on esitetty Kuvassa 10. Koostepiikkien määrän ja niiden päällekkäisyyden väheneminen oli odotettavissa kasvatettaessa ytimen leveyttä koostepiikin leveyden pysyessä vakiona (Kuvat 10A ja 10B). Väheneminen johtuu siitä, että ydinestimaatti muuttuu ylipehmentäväksi samalla yhdistäen huippupisteitä laajemmalta alueelta. Piikkimäärissä tapahtuu erikoinen hyppäys noin 27 emäsparin leveisen ytimen kohdalla. Hyppäys näkyy piikkimäärissä ja päällekkäisissä piikeissä nousuna, mutta piikittömien huippupisteiden määrässä laskuna. Hyppäyksen syystä ei ole varmuutta, mutta kuvaajien käyttäytymisestä voidaan päätellä että todennäköisesti kyseisellä leveydellä syntyy piikittömien huippupisteiden kohdalle uusia koostepiikkejä, jotka menevät toistensa tai jo olemassaolevien koostepiikkien kanssa päällekkäin.

Kääntöpuolena pelkällä ytimen levittämisellä on Kuvaajan 10C ilmiö. Kuvaajasta näkyy koostepiikkien alueelle kuulumattomien alkuperäisten huippupisteiden määrän kasvu, kun ytimen leveys kasvaa. Ilmiö johtuu koostepiikkien kapeudesta verrattuna alueeseen, jolta leveä ydin piikkejä yhdistää. Eli koostepiikkejä on vähän ja ne ovat myös kapeita, jolloin alkuperäiset huippupisteet eivät enää sijoitu koostepiikkien alueelle. Kuvaajasta 10C nähdään, että ytimen leveyteen verrattuna alle kymmenkertaiset koostepiikkien leveydet alkavat kasvattaa koostepiikittömien huippupisteiden määrää.

Seuraavaksi testattiin ytimen leveyden vaikutusta koostepiikkeihin kolmella eri ytimellä. Koostepiikin leveydeksi valittiin 10 kertaa ytimen leveys. Tulokset on esitetty Kuvassa 11. Kuvaajasta 11A voidaan nähdä, että koostepiikkien määrä pienenee ytimen leveyden kasvaessa. Kaikilla ytimen leveyksillä koostepiikkeihin kuulumattomat huippupisteet pysyvät kohtalaisen pienissä lukemissa, kasvaen hieman mentäessä yli sa-



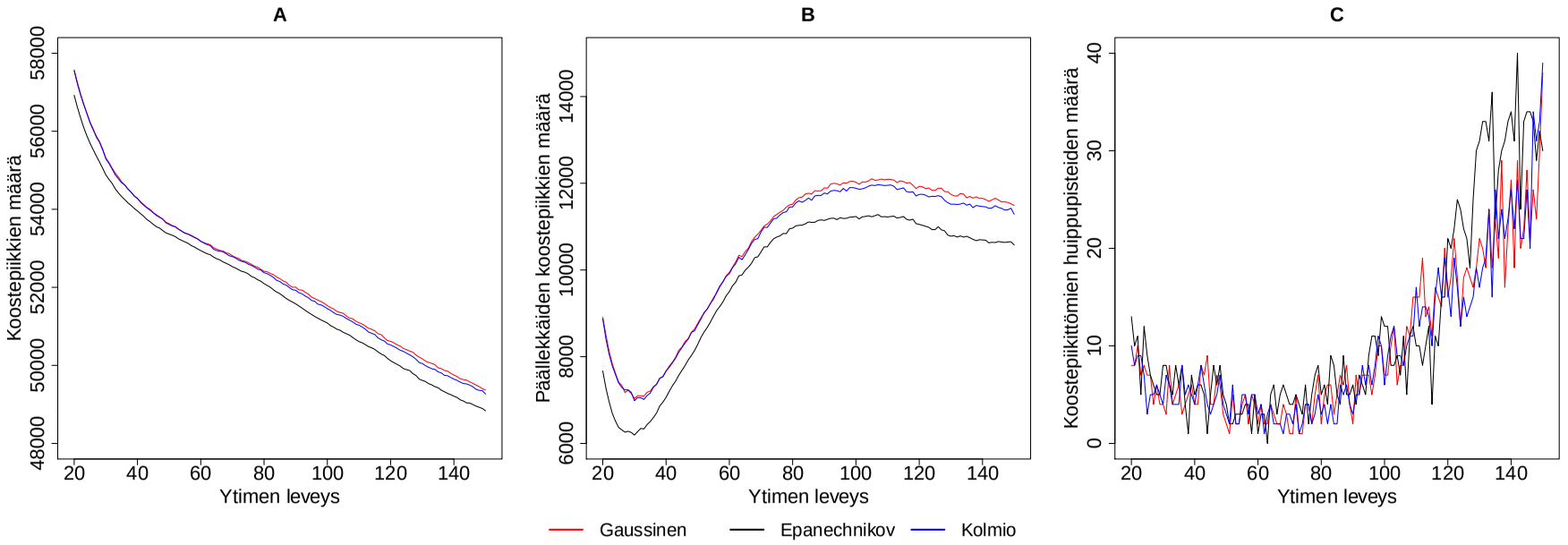

Kuva 10: Koostepiikkien (A), päällekkäisten koostepiikkien (B) ja koostepiikkeihin kuulumattomien huippupisteiden määrät (C), kun koostepiikin koko pidetään samana ytimen leveyttä kasvatettaessa. Tässä tapauksessa koostepiikin leveys oli 200 emäsparia ja ydinestimoinnissa käytettiin gaussista ydintä.

dan emäsparin levyisiin ytimiin (Kuvaaja 11C). Huippupisteiden määrä on kuitenkin moninkerroin pienempi kuin kapeamman koostepiikkileveyden tapauksessa (vrt. lukumäärät Kuva 10C vs. Kuva 11C). Päällekkäisten piikkien osalta tilanne on mielenkiintoisempi, sillä noin 30 emäsparin levyinen ydin näyttäisi tuottavan vähiten päällekkäisiä piikkejä (Kuvaaja 11B). Lisäksi samassa 30 emäsparin ytimen leveydessä myös piikkien kokonaismäärän lasku loivenee. Eri ytimien välillä ei ole suuria eroja, pois lukien Epanechnikovin ytimen pienemmät koostepiikkien ja päällekkäisten koostepiikkien määrät.

Edellisen perusteella otettiin lähempään tarkasteluun 20, 30 ja 80 levyiset ytimet. Näiden lukuarvoja on esitetty Taulukossa 9. Taulukosta selviää koostepiikkien kokonaismäärä tarkastelluilla leveyksillä gaussisella, Epanechnikovin ja kolmioytimellä. Kokonaismäärät vaihtelevat noin 52 000 koostepiikistä noin 58 000 koostepiikkiin. Lisäksi taulukkoon on laskettu myös päällekkäisten koostepiikkien määrät, vierekkäisten koostepiikkien keskietäisyydet sekä jaoteltu koostepiikit niiden sisältämien alkuperäisten huippupisteiden lukumäärien mukaan. Vaikka koostepiikittömien huippupisteiden määrä pyrittiin minimoimaan, oli jokaisessa koosteessa silti muutamia alkuperäisiä huippupisteitä, jotka eivät sijoittuneet koostepiikkien alueelle. Näiden koostepiikkeihin kuulumattomien huippupisteiden lukumäärät on ilmoitettu taulukon viimeisessä sarakkeessa.

Saadut koostepiikkien lukumäärät tarkoittavat, että parametreistä riippuen noin 20 000 - 26 000 alkuperäistä piikkiä sulautui uusiin koostepiikkeihin. Vierekkäisten koostepiikkien keskietäisyyksissä oli eroa Epanechnikovin ja kahden muun ytimen välillä joitakin satoja emäspareja riippuen ytimen leveydestä. Epanechnikovin ydin muodostaa siis samasta piikkidatasta vähemmän koostepiikkejä, jotka ovat oletetuksen mukaisesti kauempana toisistaan. Koostepiikkien etäisyyshajonta on todella suurta, joka johtuu oletettavasti alkuperäisten piikkien keskittymisestä geenien läheisyyteen. Keskittyminen tuottaa paljon lyhyitä etäisyyksiä, mutta lisää samalla pitkien etäisyyksien määrää ja mittaa.

Yhden huippupisteen sisältävien koostepiikkien lukumäärät pienenivät kaikilla ytimillä, kun ytimen leveyttä kasvatettiin. Samalla kasvoivat myös kaikki enemmän kuin yhden huippupisteen sisältävät koostepiikkiluokat. Alkuperäiset piikit sulautuivat sitä paremmin mitä leveämpi käytetty ytimen leveys on. Kääntöpuolena hyvällä sulautumisella oli kuitenkin yhä useampien koostepiikkien päällekkäisyys. Epanechnikovin ytimellä saatiin eniten suuren huippupistelukumäärän omaavia koostepiikkejä, mutta




Kuva 11: Koostepiikkien (A), päällekkäisten koostepiikkien (B) ja koostepiikkeihin kuulumattomien huippupisteiden määrät (C) eri ytimillä ytimen leveyttä kasvatettaessa. Koostepiikin leveys oli aina kymmenen kertaa käytetyn ytimen levyinen.

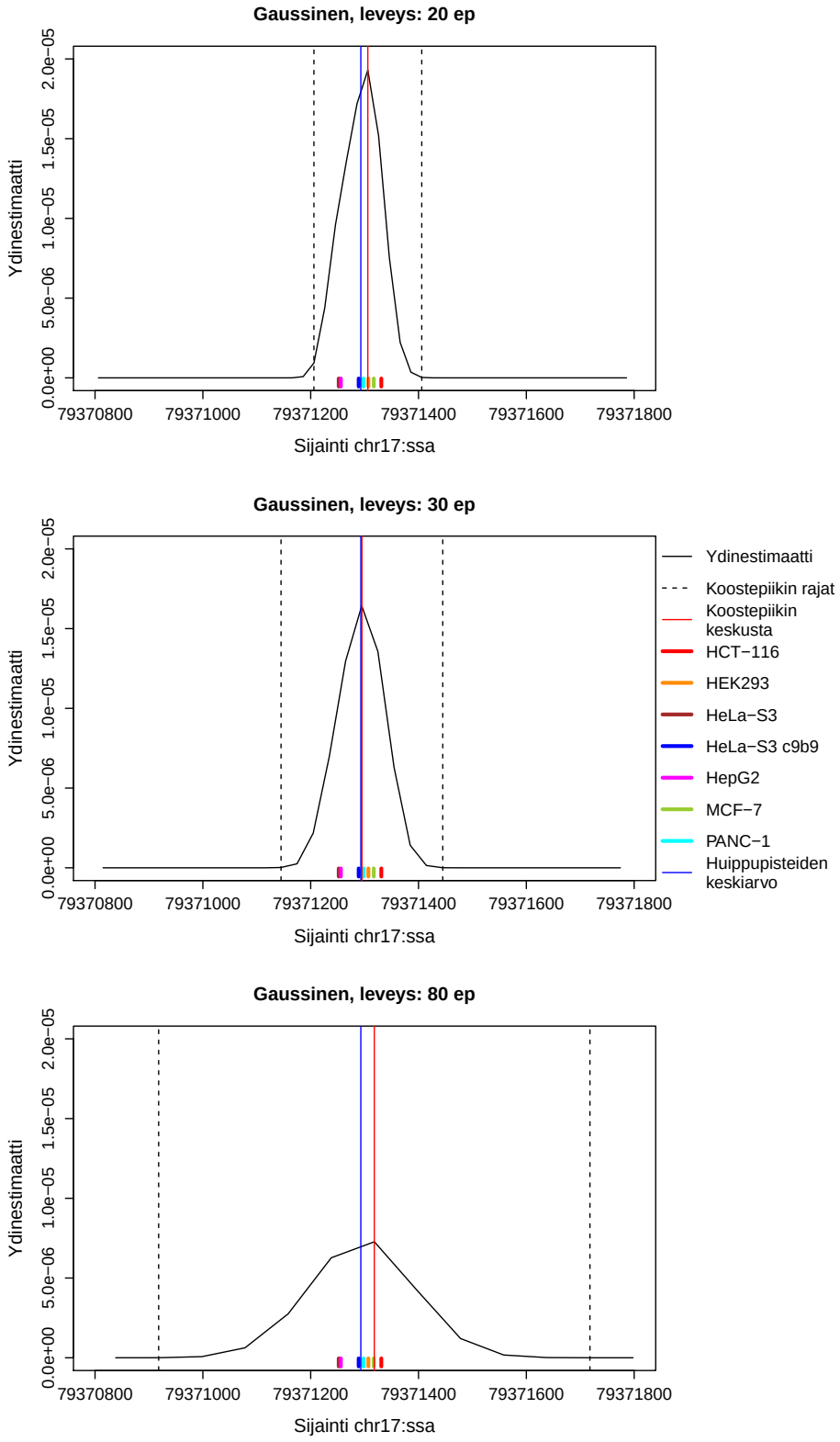
Taulukko 9: Koostepiikkitilastoja eri ytimillä ja leveyksillä. Lähtöpiikkejä oli 77874. Sarakkeet 1 hp, 2 hp, 3 hp, ... tarkoittavat sellaisten koostepiikkien lukumäärää, jossa on 1, 2, 3, ... kappaletta alkuperäisiä huippupisteitä. Sarake php kertoo niiden huippupisteiden lukumäärän, jotka eivät osuneet koostepiikkien alueelle.

| *Ydin, leveys (ep)* | *Koostepiikkejä* | *Päällekkäisiä* | *Keskietäisyys* | *≥ 7 hp* | *6 hp* | *5 hp* | *4 hp* | *3 hp* | *2 hp* | *1 hp* | *php* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gaussinen, 20 | 57574 | 8910 | $51041,3 \pm 213606,4$ | 51 | 438 | 578 | 1016 | 2696 | 7045 | 45750 | 7 |
| Gaussinen, 30 | 55318 | 7034 | $53032,1 \pm 217665,1$ | 55 | 470 | 627 | 1122 | 3172 | 7654 | 42218 | 4 |
| Gaussinen, 80 | 52411 | 11519 | $55518,95 \pm 223244,8$ | 63 | 497 | 688 | 1332 | 3740 | 8367 | 37724 | 2 |
| Epanechnikov, 20 | 56923 | 7672 | $51626,0 \pm 214754,7$ | 51 | 444 | 587 | 1031 | 2820 | 7332 | 44658 | 12 |
| Epanechnikov, 30 | 54882 | 6193 | $53454,3 \pm 218476,7$ | 54 | 472 | 637 | 1137 | 3266 | 7809 | 41507 | 8 |
| Epanechnikov, 80 | 52094 | 10965 | $55858,3 \pm 223881,0$ | 64 | 502 | 692 | 1353 | 3785 | 8480 | 37218 | 5 |
| Kolmio, 20 | 57551 | 8869 | $51061,8 \pm 213646,6$ | 51 | 443 | 581 | 1015 | 2753 | 6918 | 45790 | 9 |
| Kolmio, 30 | 55291 | 6986 | $53058,2 \pm 217715,1$ | 55 | 470 | 630 | 1129 | 3206 | 7580 | 42221 | 4 |
| Kolmio, 80 | 52376 | 11453 | $55556,4 \pm 223314,7$ | 63 | 499 | 692 | 1342 | 3748 | 8329 | 37703 | 3 |

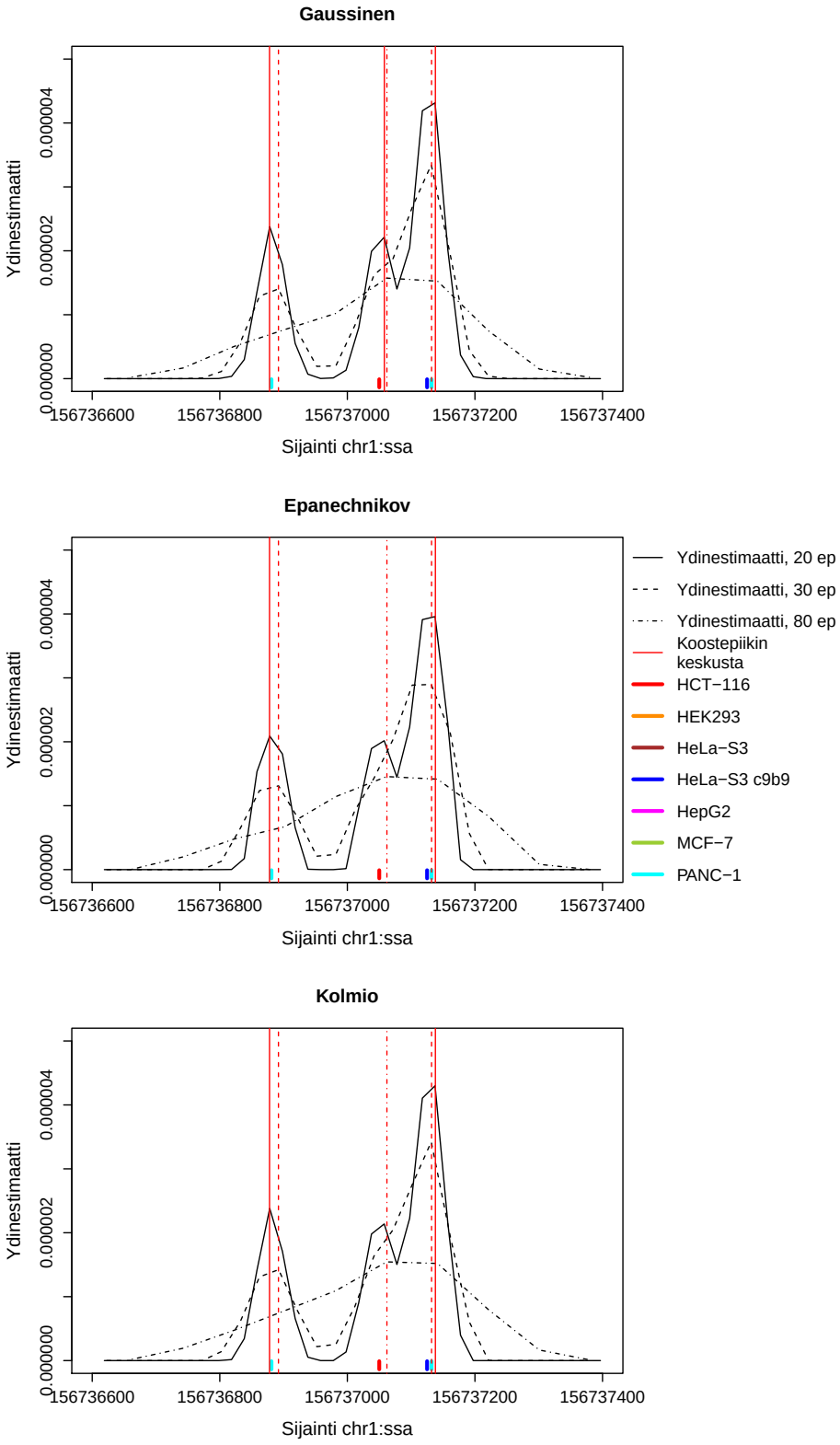
Epanechnikovin ytimellä oli myös suurimmat määrät piikittömiä huippupisteitä.

Taulukon 9 lukujen eroja eri ytimien ja ytimien leveyksien välillä tarkasteltiin kahden datasta poimitun, paljon toisistaan poikkeavan, alueen ydinestimaatteja. Ensimmäinen tarkasteltava alue on yksinkertainen, mutta harvinainen ideaalitapaus, missä monen eri solulinjan huippupisteet ovat lähekkäin (Kuva 12). Kuvassa on merkitty eri solulinjojen huippupisteet kuvaajan alareunaan eri väreillä. Pitkät pystyviivat kertovat muodostuneen koostepiikin rajat (musta katkoviiva) ja koostepiikin keskustan eli estimaatin huippukohdan (punainen pystyviiva). Lisäksi kuvaan on merkitty koostepiikkiin kuuluvien huippupisteiden keskiarvo (sininen pystyviiva), jonka tällaisessa selkeässä tilanteessa voi olettaa olevan mahdollisimman lähellä koostepiikin keskustaa. Kuvassa on esitetty vain gaussisen ytimen tulokset, koska tällaisissa tapauksissa eri ytimet eivät tarkastelujen perusteella juuri eroa toisistaan ja ytimen leveyskin vaikuttaa vain koostepiikin huipun sijaintiin. Taulukkoon 9 Kuvan 12 kaltainen tapaus ei siis juurikaan vaikuta.

Suurimmat muutokset Taulukossa 9 saa aikaan toisena tarkasteltu alue (Kuva 13). Kuvan tapaus on ongelmallinen sekä valitettavan yleinen. Kuvassa eri ytimen leveyksillä muodostetut estimaatit (mustalla) ja niiden huippukohdat (punaiset pystyviivat) on piirretty jokaiselle eri ytimen leveydelle erinäköisillä viivoilla. Kapein 20 emäsparin estimaatti ja sen keskikohdat on piirretty yhtenäisellä viivalla, 30 emäsparin levyinen estimaatti keskikohtineen katkonaisella viivalla ja 80 emäsparin levyinen estimaatti keskikohtineen katkonaisella pisteviivalla. Kuvan tapauksessa on kolme toisistaan hieman erillään olevaa huippupistekeskittymää, minkä vuoksi eri levyisten ytimien tuottamat estimaatit eroavat toisistaan. Tämän lisäksi eroa löytyy myös eri ytimien välillä. Oikealla kuvassa olevat kaksi rykelmää muodostavat 20 emäsparin levyisellä gaussisella ytimellä kaksi koostepiikkiä, mutta 30 emäsparin levyisellä ytimellä vain yhden. Epanechnikovin ytimellä ja kolmioytimellä oikeanpuoleisista rykelmistä muodostuu vain yksi koostepiikki jo heti 20 emäsparin levyisellä ytimellä. Kaikilla ytimillä 80 emäsparin leveys tuottaa vain yhden koostepiikin. Kuvan 13 kaltaisten alueiden vuoksi parametrien valinnassa joudutaan aina miettimään rajoja sille, kuinka lähekkäin olevat huippupisteet halutaan vielä omiksi koostepiikeikseen.




Kuva 12: Esimerkki 20 ep, 30 ep ja 80 ep levyisen gaussisen ytimen vaikutuksesta koostepiikkiin harvinaisessa ideaalitapauksessa, jossa alkuperäiset piikkihuiput monesta solulinjasta ovat lähekkäin. Epanechnikov ja kolmio -ytimet eivät tässä tapauksessa eronneet yhtään gaussisesta ytimestä.




Kuva 13: Esimerkki yleisestä ongelmatapauksesta, jossa eri ytimillä ja ytimen leveyksillä 20, 30 ja 80 saadaan eri määrä koostepiikkejä. Eri leveyksillä muodostettujen koostepiikkien keskustat on esitetty samanlaisella, mutta punaisella, viivalla kuin sen estimaatti. Kuvan alueella ei ole jokaisen solulinjan huippupistettä.

### 5.3.2 Sitoutumismotiivit ja yhdenmukaisuus ENCODE:n tulosten kanssa

TCF7L2-motiivin esiintyvyyttä koostepiikeissä tutkittiin samoilla ytimillä ja ytimen leveyksillä kuin Taulukossa 9. Tulokset motiivin esiintyvyydelle koostepiikeissä ja motiivin etäisyydelle koostepiikin huippupisteestä on ilmoitettu Taulukossa 10. Motiiveja esiintyi kaikilla tutkituilla ytimillä ja ytimen leveyksillä yli 20 %:ssa piikkejä. Koostepiikkien alueelle sattuvien motiivien määrä kasvoi ytimen ja koostepiikin leventyessä, mutta samalla myös kasvoi motiivin keskietäisyys koostepiikin huippupisteestä. Käytetyllä ydinfunktiolla ei näyttänyt olevan juurikaan merkitystä motiivien esiintyvyyteen ja etäisyyteen koostepiikin keskustasta.

Taulukko 10: Motiivipiikkien osuus koostepiikeistä sekä motiivien keskietäisyys koostepiikin keskustasta kolmella ydinfunktiolla ja ytimen leveydellä.

| *Ydin, leveys (ep)* | *Motiivipiikkejä (% koostepiikeistä)* | *Motiivin keskietäisyys (ep)* |
|---|---|---|
| Gaussinen, 20 | 11703 (20,3 %) | $0,3 \pm 35,7$ |
| Gaussinen, 30 | 12543 (22,7 %) | $2,2 \pm 48,8$ |
| Gaussinen, 80 | 15376 (29,3 %) | $14,5 \pm 142,5$ |
| Epanechnikov, 20 | 11694 (20,5 %) | $0,2 \pm 35,6$ |
| Epanechnikov, 30 | 12531 (22,8 %) | $2,1 \pm 48,8$ |
| Epanechnikov, 80 | 15360 (29,5 %) | $14,2 \pm 143,1$ |
| Kolmio, 20 | 11706 (20,3 %) | $0,3 \pm 35,7$ |
| Kolmio, 30 | 12539 (22,7 %) | $2,1 \pm 48,9$ |
| Kolmio, 80 | 15372 (29,3 %) | $14,5 \pm 142,6$ |

Solulinjakohtaiset motiiviesiintyvyydet koostamiseen käytetyissä alkuperäisissä piikeissä on esitetty Taulukossa 11. Verrattaessa koostepiikkien motiivietäisyyksiä alkuperäisten piikkien motiivietäisyyksiin, ovat 80 emäsparin levyiset ytimet tutkituista ytimistä ainoita, joilla motiivit ovat keskimäärin kauempana keskustasta kuin alkuperäisillä piikeillä. Motiivin sisältävien koostepiikkien lukumäärät vaihtelevat 11694:stä 15376:een. Motiivin sisältävien alkuperäisten piikkien määrä puolestaan vaihtelee 1112:sta 6233:een. Suurin absoluuttinen määrä motiivipiikkejä oli solulinjassa HCT-116 ja pienin määrä solulinjassa HeLa-S3. Suhteutettuna piikkien kokonaismäärään oli HCT-116:ssa kuitenkin vähiten motiivipiikkejä (21,3 %) ja solulinjassa HepG2 eniten (65,0 %, 1122 motiivipiikkiä).

Motiivien esiintyvyyksien vertailu yksittäisten näytteiden ja koostepiikkien kesken ei ole kuitenkaan järkevää. Todennäköisesti ainakin osa alkuperäisistä motiivipiikeistä on samoja eri solulinjoissa ja ne näkyvät koostepiikkien tapauksessa vain yhtenä motiivin

sisältävänä koostepiikkinä. Vaihtoehto on verrata esimerkiksi alkuperäisten piikkien kokonaismääriä. Alkuperäiset piikit yhteenlaskettuna (77874 kpl) ja alkuperäiset motiivipiikit yhteenlaskettuna (21352 kpl) saadaan noin 27,4 % motiiviesiintyvyys alkuperäisissä piikeissä. Tätä lähimpänä ovat 80 levyisten ytimien koostepiikit (ero 1,9 - 2,1 prosenttiyksikköä) ja seuraavana 30 levyisten ytimien koostepiikit (ero 4,7 - 4,6 prosenttiyksikköä). Alkuperäisten piikkien yhteenlasketuissa lukumäärissä ei tosin myöskään näy tietoa siitä, kuinka monessa piikissä on kyse samasta motiivin esiintymispaikasta.

Taulukko 11: Motiivipiikkien osuus alkuperäisistä piikeistä sekä motiivien keskietäisyys piikin huippupisteestä alkuperäisissä piikkidatoissa.

| *Solulinja* | *Motiivipiikkejä (% piikeistä)* | *Motiivin keskietäisyys (ep)* |
|---|---|---|
| HepG2 | 1122 (65,0 %) | $13,0 \pm 90,4$ |
| HCT-116 | 6233 (21,3 %) | $4,6 \pm 103,4$ |
| HEK293 | 2887 (37,4 %) | $5,6 \pm 92,9$ |
| HeLa-S3 | 1112 (45,7 %) | $4,0 \pm 71,6$ |
| HeLa-S3 c9b9 | 4025 (23,7 %) | $5,6 \pm 102,9$ |
| PANC-1 | 3269 (27,3 %) | $11,6 \pm 115,8$ |
| MCF-7 | 2704 (34,8 %) | $3,8 \pm 73,9$ |

Tähän analyysiin käytetyistä ENCODE:n TCF7L2:n datoista on myös ENCODE-projektissa muodostettu TCF7L2:n yhteiset sitoutumisalueet. ENCODE:n sitoutumisalueita verrattiin 30 emäsparin levyisen Epanechnikovin ytimen tuottamiin koostepiikkeihin. Epanechnikovin ydin valittiin, koska se vaikutti koostavan piikkejä parhaiten.

ENCODE:n määrittämiä TCF7L2-sitoutumisalueita on 45100. Tämä on huomattavasti vähemmän kuin minkään tässä käytetyn ytimen tuottama määrä. ENCODE:n sitoutumisalueiden leveyskin tosin vaihtelee suuresti: pienimmillään ne ovat 94 emäsparia leveitä ja suurimmillaan 1853 emäsparia leveitä alueita (keskiarvo 530,2 $\pm$ 134,6). Silmämääräisesti tutkittuna ENCODE:n sitoutumispaikat on muodostettu siten, että kaikki lähellä toisiaan olevat piikit on yhdistetty yhdeksi alueeksi. Näin ollen ENCODE:n tapa yhdistää piikit tuottaa lähinnä yleisiä alueita, joille TCF7L2 sitoutuu, erottelematta tarkasti eri solulinjojen sitoutumisalueita.

Eroavaisuudet ENCODE:n alueiden ja 30 emäsparin levyisen Epanechnikovin ytimen tuottamien koostepiikkien välillä laskettiin sekä emäspari että piikkitasolla käyttäen Kaavoja 1 ja 2. Emäsparitasolla $\alpha$ oli 50,6 % ja $\beta$ oli 26,7 %. ENCODE:n sitoutumisa-

lueista noin puolet sattuu siis alueille, joilla ei ole koostepiikkejä. Koostepiikeistä puolestaan vain neljäsosa on alueilla, jotka eivät ole ENCODE:n määrittämiä sitoutumisalueita. Alue-/koostepiikkitasolla $\alpha$ oli 15,9 % ja $\beta$ oli 23,2 %. Koostepiikkejä oli siis suurempi osa alueilla, joissa ei ollut ENCODE:n aluetta, kuin ENCODE:n sitoutumisalueita alueilla, joissa ei ollut koostepiikkejä. Taulukossa 12 on esitetty ENCODE:n sitoutumisalueilla sijaitsevien koostepiikkien lukumäärät ja päinvastoin. ENCODE:n sitoutumisalueilla sijaitsi joissakin tapauksissa useampikin koostepiikki. Koostepiikkien osalta puolestaan oli harvinaista, jos koostepiikin alueella oli useampi kuin yksi ENCODE:n sitoutumisalue.

Koska ENCODE:n sitoutumisalueet olivat keskimäärin leveämpiä kuin koostepiikit, laskettiin miten paljon ENCODE:n sitoutumisalueiden emäspareista oli piikeissä, jotka oli koostepiikkien kanssa päällekkäin ja miten paljon ei. 3733900 emäsparia eli noin 15,6 % kaikista ENCODE:n alueilla sijaitsi koostepiikittömillä sitoutumisalueilla. Kun ottaa huomioon aiemmin saadun 50,6 %, tarkoittaa se että loput 23,8 % emäspareista sijaitsee koostepiikkien alueella. Koostepiikit ovat siis huomattavasti tarkempia alueita kuin ENCODE:n piikit, koska alue/koostepiikkitasolla ENCODE:n alueista 84,1 % on mukana koostepiikeissä, mutta nukleotiditasolla vain 23,8 %.

Taulukko 12: Koostepiikkien päällekkäisyys ENCODE:n sitoutumisalueiden kanssa. Tapaus 1: ENCODE:n sitoutumisalueilla sijaitsevien koostepiikkien lukumäärät. Tapaus 2: Koostepiikkien alueella sijaitsevien ENCODE:n sitoutumisalueiden lukumäärät.

| *Tapaus* | *0 kpl* | *1 kpl* | *2 kpl* | *3 kpl* | *4 kpl* | *5 kpl* |
|---|---|---|---|---|---|---|
| Tapaus 1 | 7189 | 34063 | 3431 | 377 | 37 | 3 |
| Tapaus 2 | 12722 | 42101 | 59 | 0 | 0 | 0 |

## 5.4 Tulosten tulkinta

Menetelmälle sopivimpien parametrien selvityksessä parhaimmaksi osoittautui 30 emäsparin levyinen Epanechnikovin ydin. Kyseinen ydin antaa vähiten sekä koostepiikkejä että päällekkäisiä koostepiikkejä. Koostamiskyky 30 emäsparin levyisellä Epanechnikovin ytimellä on parempi kuin muilla ytimillä, koska yhden huippupisteen koostepiikkejä on vähemmän ja monen huippupisteen koostepiikkejä enemmän muihin ytimiin verrattuna. Toisin sanoen saadaan pienemmät koostepiikkimäärät menettämättä yhtään enempää alkuperäisiä huippupisteitä kuin muutkaan tässä tutkitut ytimet. Paras

koostepiikin leveys näyttäisi olevan kymmenkertainen ytimen leveys, jolla pystytään minimoimaan päällekkäisten koostepiikkien määrää tuloksissa.

Testauksen tuloksena saadut parametrit eivät saata kuitenkaan soveltua kaikkiin mahdollisiin datoihin, mutta niitä voi pitää hyvänä lähtökohtana soveltuvimpien parametrien etsimiseksi. Tässä saatujen tulosten perusteella tärkeintä on kokeilla etenkin eri levyisiä ytimiä, kun menetelmää käytetään uudelle datalle. Tulosten perusteella käytetyllä ytimellä ei ole niin suurta merkitystä, mutta Epanechnikovin ydin suoriutuu parhaiten.

Menetelmä suoriutui hyvin koostepiikkien keskittämisessä motiiveihin paitsi leveiden ytimien (> 80 ep) tapauksessa. Verrattuna alkuperäisiin piikkeihin motiivin etäisyys piikin keskustasta pieneni hiukan. Suurin vaikutus ytimen leveydellä oli motiivin etäisyyden hajontaan, joka pieneni huomattavasti testattujen kapeampien ytimien tapauksessa verrattuna alkuperäisiin piikkeihin. Motiivipiikkien määrä oli prosentuaalisesti pienempi kuin alkuperäisten piikkien tapauksessa, mikä voi johtua siitä että monella eri solulinjalla on todennäköisesti saman motiivin kohdalla piikki. Vallalla olevan käsityksen mukaan suurin osa sitoutumisesta pitäisi tapahtua sitoutumismotiivien kohdalla, joten koostepiikeissä voisi olettaa olevan enemmän huippupisteitä etenkin motiivien kohdalla.

Koostepiikkien ja ENCODE:n sitoutumisalueiden vertailussa tulokset eivät olleet ratkaisevia. Koostepiikit olivat suurimmassa osassa tapauksia ENCODE:n sitoutumisalueita kapeampia. Isossa osassa ENCODE-sitoutumisalueita oli myös useampi kuin yksi koostepiikki. Koostepiikit pystyivät useammassa tapauksessa tarkentamaan ENCODE:n sitoutumisalueiden sijaintia, koska nukleotiditasolla ENCODE:n sitoutumisalueita oli enemmän kuin koostepiikkejä, mutta piikkitasolla toisinpäin. Sekä ENCODE:n sitoutumisalueissa että koostepiikeissä oli alueita/piikkejä, joita toisella ei ollut. Johtopäätöksenä voidaan pitää, että ENCODE:n sitoutumisalueet kertovat yleisestä sitoutumispaikasta ja koostepiikit kertovat sitoutumispaikat tarkemmin. Tarkempaa kuin yleistä sitoutumispaikkatietoa tarvittaessa on siis järkevämpää muodostaa koostepiikit kuin käyttää ENCODE:n sitoutumisalueita.

Testausta tehdessä tuli esille muutamia kehityskohteita, joiden lisätutkiminen voisi olla hyödyllistä. Ensimmäinen kehityskohde on koostepiikkien leveyden määritys, vaikka testauksessa sille hyvä ohjeellinen arvo löydettiinkin. Parempi vaihtoehto leveyden määrittämiseksi voisi olla johtaa koostepiikkien leveys jotenkin alkuperäisten piikkien

leveydestä, esimerkiksi leveyksien keskiarvosta. Toisena mahdollisena kehityskohteena voisi olla painotettu ydinestimointi, jossa tietyille piikkien huippupisteille voisi antaa enemmän painoarvoa. Painoarvo voisi esimerkiksi perustua joko piikkisignaalin voimakkuuteen tai piikin tilastolliseen merkitsevyyteen. Painotettu ydinestimointi saattaisi tuoda selkeyttä esimerkiksi Kuvan 13 tyyppisiin tapauksiin, missä ydinestimaatin keskittyminen aina varmimpien piikkien huippupisteisiin vähentäisi eri parametrien vaikutusta.

# 6 Johtopäätökset

Eri tutkimuksissa saatujen ChIP-seq-kokeiden tulosten hyödyntäminen on muuttunut yhä houkuttelevammaksi saatavilla olevan datamäärän kasvaessa. Yksi lähestymistapa usean ChIP-seq-kokeen hyödyntämiseen on ChIP-seq-piikkien koostaminen, jossa usean eri ChIP-seq-kokeen piikit yhdistetään koosteeksi. Yhtenä vaihtoehtona koostamiselle on ydinestimointi, jota soveltamalla voidaan käyttää hyväksi piikkien tiheyttä. Sekä ChIP-seq:stä että ydinestimoinnista löytyy paljon kirjallisuutta, johon tutustumalla tässä tutkimuksessa kehitettiin ydinestimointiin perustuva ChIP-seq-piikkien koostamismenetelmä.

ChIP-seq-kokeiden yhdistämisessä on omat haasteensa. ChIP-seq-menetelmässä itsessään on monia muuttujia, jotka saattavat tehdä eri aikaan ja eri tutkijoiden tekemistä kokeista vaihtelevia ja vaikeasti verrattavia. Lisäksi ChIP-seq-datan käsittelyyn on kehitetty monia erilaisia työkaluja, jotka tuovat omat muuttujansa saatuihin tuloksiin. Näistä ongelmista jälkimmäiseen voi vaikuttaa analysoimalla datat uudestaan käyttäen samoja työkaluja, mutta ensimmäiseen ongelmaan ei ole olemassa helppoja ratkaisuja. Paras tapa on tutustua eri virheenlähteisiin ja ottaa ne huomioon tuloksia tulkittaessa.

Ydinestimointia on ChIP-seq:n yhteydessä tähän mennessä käytetty tiettävästi vain ChIP-seq-piikkien hakuun ChIP-seq-datasta. Ydinestimoinnissa on kaksi tärkeää parametria: ytimen muoto ja ytimen leveys. Parametreista jälkimmäinen on tuottanut vaikeuksia genomisella ChIP-seq-datalla, vaikka ytimen leveyden määrittämiseksi on kehitetty automaattisia menetelmiä. Automaattisia leveydenmääritysmenetelmiä ei nimittäin ole kehitetty silmälläpitäen erittäin laajalle alalle jakautunutta dataa, jollaista kaikki genominen data on. Tästä syystä ChIP-seq-datalle tehdyissä menetelmissä on päädytty joko käyttäjän antamaan kiinteään leveyteen tai mukautuvan leveyden ydinestimointiin.

Tässä tutkimuksessa esitettiin tiettävästi ensimmäinen, varsinainen ChIP-seq-piikkien koostamiseen tarkoitettuja menetelmä: ConsensusSummit. ConsensusSummit-menetelmä perustuu kiinteän leveyden ydinestimointiin. Menetelmän syötteenä toimivat ChIP-seq-piikkien huippupisteet ja se antaa tulosteena koostepiikit, joissa samassa sijainnissa olevat eri ChIP-seq-kokeiden piikit ovat esitettynä yhtenä koostepiikkinä. Koostepiikit muodostetaan käyttäen apuna ChIP-seq-piikkien huippupisteiden tiheyttä, joka arvioidaan ydinestimointia käyttäen. Ensimmäinen versio

ConsensusSummit-menetelmästä kehitettiin D-vitamiinireseptorin tutkimista varten, mutta menetelmää on sen jälkeen kehitetty lisää.

ConsensusSummit-menetelmän testaus on jäänyt vähäiselle huomiolle, minkä vuoksi menetelmää testattiin ENCODE:n TCF7L2-tranksriptiofaktorin datalla. Testausaineisto valittiin, koska sille on olemassa vertailukelpoinen koosteenkaltainen kokoelma sitoutumisalueita. Testauksen ohessa etsittiin myös menetelmälle sopivia parametreja. Sopivimmaksi parametriksi menetelmälle todettiin 30 emäsparin levyinen Epanechnikovin ydin. Parametrien toimivuutta on kuitenkin syytä aina hieman tarkastella uusia datoja koostettaessa. Testauksen perusteella ConsensusSummit-menetelmää sopii hyvin ChIP-seq-piikkien koostamiseen. ENCODE:n vertailuaineistoon verrattuna ConsensusSummit-menetelmän koostepiikit ovat yleisesti ottaen kapeampia. ConsensusSummit-menetelmä myös erottelee yksittäisiä ENCODE:n vertailuaineiston sitoutumisalueita kapeammiksi ja tarkemmiksi sitoutumisalueiksi. Erityisen hyvin ConsensusSummit-menetelmä koostaa piikkejä, jotka ovat keskittyneet sitoutumismotiivien läheisyyteen.

Saatujen tulosten perusteella ChIP-seq-piikkien koostaminen on mahdollista käyttäen ydinestimointia. Kehitetty ConsensusSummit-menetelmä on hyödyllinen työkalu, jolla saadaan uudenlaista tietoa ChIP-seq-piikkien sijoittumisesta genomiin. Menetelmän avulla voidaan tutkia myös esimerkiksi sitoutumismotiiveja ja menetelmää voi olla mahdollista hyödyntää myös esimerkiksi aikasarja-perusteisissa ChIP-seq-tutkimuksissa, joissa eri aikapisteiden piikkien samankaltaistaminen helpottaa jatkoanalyyseja.

Menetelmässä olisi kuitenkin varaa myös jatkokehitykselle. Paras jatkokehityksen kohde olisi laajalle levittäytyneelle datalle suunniteltu automaattinen leveydenmääritysmenetelmä. Automaattisen leveydenmääritysmenetelmän kehittäminen voisi johtaa ydinestimoinnin käytön yleistymiseen myös muussakin genominlaajuisessa tutkimuksessa. ConsensusSummit-menetelmän kannalta mielenkiintoinen tutkimuksen kohde voisi olla myös jonkinlainen painoitettu ChIP-seq-piikkien ydinestimointi, jossa alkuperäisistä piikeistä käytetään muitakin arvoja kuin pelkkää huippupistettä, esimerkiksi piikkien tilastollista merkitsevyyttä. Painotuksen avulla menetelmälle saattaisi löytyä uusia käyttökohteita.

# Viitteet


Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, ja Jie Zhang. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Computational Biology*, 9(11): e1003326, 2013. doi: 10.1371/journal.pcbi.1003326.

Alan P. Boyle, Justin Guinney, Gregory E. Crawford, ja Terrence S. Furey. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21): 2537–2538, 2008.

Li Chen, Chi Wang, Zhaohui S. Qin, ja Hao Wu. A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics*, 2015. doi: 10.1093/bioinformatics/btv094.

Maria D. Chikina ja Olga G. Troyanskaya. An effective statistical evaluation of ChIP-seq dataset similarity. *Bioinformatics*, 28(5):607–613, 2012.

Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, ja Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

Peggy J. Farnham. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, 10(9):605–616, 2009.

Melissa J. Fullwood, Chia-Lin Wei, Edison T. Liu, ja Yijun Ruan. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research*, 19(4):521–532, 2009.

Terrence S. Furey. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, ja Christopher K.

Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4):576–589, 2010.

David S. Johnson, Ali Mortazavi, Richard M. Myers, ja Barbara Wold. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502, 2007.

M. C. Jones, J. S. Marron, ja Simon J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433): 401–407, 1996.

Peter V. Kharchenko, Michael Y. Tolstorukov, ja Peter J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12): 1351–1359, 2008.

Jan O. Korbel, Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, Dean Palejev, Nicholas J. Carriero, Lei Du, Bruce E. Taillon, Zhoutao Chen, Andrea Tanzer, A. C. Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P. Carter, Matthew E. Hurles, Sherman M. Weissman, Timothy T. Harkins, Mark B. Gerstein, Michael Egholm, ja Michael Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, Peter Bickel, James B. Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Jason Gertz, Alexander J. Hartemink, Michael M. Hoffman, Iyer R. Vishwanath, Youngsook L. Jung, Subhradip Karmakar, Kellis Manolis, Peter V. Kharchenko, Qunhua Li, Tao Liu, X. Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M. Myers, Peter J. Park, Michael J. Pazin, Marc D. Perry, Debasish Raha, Timothy E. Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A. Stamatoyannopoulos, Michael Y. Tolstorukov, Kevin P. White, Simon Xi, Peggy J. Farnham, Jason D. Lieb, Barbara J. Wold, ja Michael Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, 2012.

Ben Langmead, Cole Trapnell, Mihai Pop, ja Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009. doi: 10.1186/gb-2009-10-3-r25.

Heng Li ja Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

Edison T. Liu, Sebastian Pott, ja Mikael Huss. Q&a: Chip-seq technologies and the study of gene regulation. *BMC Biology*, 8(1):56, 2010.

Kazumitsu Maehara, Jun Odawara, Akihito Harada, Tomohiko Yoshimi, Koji Nagao, Chikashi Obuse, Koichi Akashi, Taro Tachibana, Toshio Sakata, ja Yasuyuki Ohkawa. A co-localization model of paired ChIP-seq data using a large ENCODE data set enables comparison of multiple samples. *Nucleic Acids Research*, 2012. doi: 10.1093/nar/gks1010.

Michael L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

Peter J. Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Shirley Pepke, Barbara Wold, ja Ali Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6:S22–S32, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

Parameswaran Ramachandran ja Theodore J. Perkins. Adaptive bandwidth kernel density estimation for next-generation sequencing data. In Victor Jin, editor, *BMC Proceedings*, volume 7, page S7. BioMed Central Ltd., 2013.

Vikas C. Raykar ja Ramani Duraiswami. Fast optimal bandwidth selection for kernel density estimation. In Joydeep Ghosh, Diane Lambert, David Skillicorn, ja Arnold Goodman, editors, *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 524–528. Society for Industrial and Applied Mathematics., 2006.

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, ja Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Rimantas Rudzkis ja Mindaugas Kavaliauskas. On local bandwidth selection for density estimation. *Informatica*, 9(4):479–490, 1998.

Stephan R. Sain ja David W. Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 91(436):1525–1534, 1996.

Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Evan Bolton, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M. Fingerman, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Zhiyoung Lu, Thomas L. Madden, Tom Madej, Donna R. Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrachi, James Ostell, Anna Panchenko, Lon Phan, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Yanli Wang, W. John Wilbur, Eugene Yaschenko, ja Jian Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(suppl 1): D38–D51, 2011.

Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46 (1):561–584, 1995.

Zhen Shao, Yijing Zhang, Guo-Cheng Yuan, Stuart H. Orkin, ja David J. Waxman. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biology*, 13(3):R16, 2012.

Simon J. Sheather. Density estimation. *Statistical Science*, 19(4):588–597, 2004.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, Lontoo, 1986.

Adi L. Tarca, Vincent J. Carey, Xue-wen Chen, Roberto Romero, ja Sorin Drăghici. Machine learning and its applications to biology. *PLoS Computational Biology*, 3 (6):e116, 2007. doi: 10.1371/journal.pcbi.0030116.

George R. Terrell. The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85(410):470–477, 1990.

George R. Terrell ja David W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.

Pauli Tuoresmäki, Sami Väisänen, Antonio Neme, Sami Heikkinen, ja Carsten Carlberg. Patterns of genome-wide VDR locations. *PloS ONE*, 9(4):e96105, 2014. doi: 10.1371/journal.pone.0096105.

Anton Valouev, David S. Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M. Myers, ja Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834, 2008.

Nava Whiteford, Niall Haslam, Gerald Weber, Adam Prügel-Bennett, Jonathan W. Essex, Peter L. Roach, Mark Bradley, ja Cameron Neylon. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, 33(19):e171, 2005. doi: 10.1093/nar/gni170.

Elizabeth G. Wilbanks ja Marc T. Facciotti. Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS ONE*, 5(7):e11471, 2010. doi: 10.1371/journal.pone.0011471.

Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, ja X. Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9 (9):R137, 2008. doi: 10.1186/gb-2008-9-9-r137.

# Liite 1: ChIP-seq tietokantoja

ChIP-seq-datalle on useita julkisia tietokantoja, joista kaksi tärkeintä ovat Euroopan ja Yhdysvaltojen kansallisten instituutioiden tietokannat.

Eurooppa:
European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA)
Osoite: http://www.ebi.ac.uk/ena

Yhdysvallat:
National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)
Julkaisu: Sayers ym. 2011
Osoite: http://www.ncbi.nlm.nih.gov/Traces/sra2