Genetika, zelulen, molekulen eta eboluzioaren biologiaren esparru barneko esperimentazioaren hastapena

5 Gaia. Populazio analisien Teknikak

ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

NAZIOARTEKO BIKAINTASUN CAMPUSA
CAMPUS DE EXCELENCIA INTERNACIONAL

Dibertsitate Genetikoaren Kontzeptua

- Dibertsitate genetikoa, espezie bakoitzean existitzen diren **ezaugarri genetikoen kopuru** totala da. Dibertsitate genetikoa garrantzitsua da, hautespen naturalak eragiten dion lehengaia delako: dibertsitate ezean, ez dago eboluziorik.
- **Dibertsitate kantitatea** ondoko parametroen bidez estimatu daiteke:
  - mutazio-tasa (dibertsitatearen sortzailea)
  - populazioaren tamaina (honi lotuta, jito genetikoa)
  - migrazioa eta hautespena
  - eta neurri batean, errekonbinazioa, nahiz eta dibertsitate berririk ez sortu, existitzen dena berrantolatzen du.

• Demagun:

p_1, A-ren maiztasuna analizatu den populazioko laginean
p_2, B-ren maiztasuna
($p_1 + p_2 = 1$ delarik)

• Hardy-Weinberg orekan, homozigotoen esperotako maiztasuna, heuren maiztasun alelikoen karratua da, beraz:

esperotako Homozigositatea $p_1^2 + p_2^2$ da

hortaz, esperotako Heterozigositatea (dibertsitate genetikoa) zera da:

1 –Homozigositatea, beraz:

Esperotako heterozigositatea= $1 - \sum_{i=1}^{n} p_i^2$

Edo, zuzenagoa dena: $n*(1-\sum_{i=1}^{n} p_i^2)/(n-1)$

n, alelo kopurua izanik, kasu honetan 2 da.

DNA sekuentzia eta haplotipo kontzeptua

- **DNA sekuentziak**, markari genetiko klasikoak eta agian SNP independenteak baino askoz **informagarriagoak** dira.
- Eskualde genomiko zehatz bat sekuentziatu ondoren, aita eta amarengandik jasotako genomen informazio nahastua dugu eta teknika esperimentalak edo/eta bioinformatikoak erabili behar dira bi informaziook banatu eta **kromosoma berean** zein **aldaki aleliko** doazen jakiteko (adb. Haplotipoak edo fase haplotipikoak lortu).
- Genoma mitokondrial edo Y-kromosomako eskualde pseudoautosomikoan gertatzen den ez bezala, ***locus* autosomiko** batetan berdinak zein ezberdinak izan daitezkeen **bi haplotipo** agertuko dira.
- Haplotipoen informagarritasun handiena, sekuentzia haplotipikoaren informazioak (aldaki aleliko lotu sorta), haplotipo **zaharrenen** eta **gazteenen** artean ezberdintzatzea eta zein haplotipo eratortzen den beste haplotipo batetatik jakitea baimentzen digula da.
- Hots, analisi haplotipikoari esker, *locus* edo eskualde genomiko berean behatu izan diren haplotipo desberdinen **genealogia ondorioztatu** dezakegu.

Dibertsitatea DNA sekuentzietan

- DNA sekuentzien kasuan, **heterozigositateari** buruz hitz-egin beharrean, informagarriagoa da **dibertsitate genetikoari** buruz hitz-egitea.
- **Teoria Neutrala** kontutan izanik, maila molekularrean, **aldaketa** edo mutazio **gehienak neutroak** direla onartu behar da, abantaila zein desabantaila gabe. Honela, oreka batetara heltzen gara, zeinetan mutazioz sortzen den aldakortasun kantitate berria, populazio tamaina mugatua izateagatik, jito genetikoz, galtzen denarekin orekatzen den.
- Hala, orekan, esperotako dibertsitate teorikoa (θ), zera da:

$$\theta = 4N_e\mu$$

N_e, populazioaren tamaina efektiboa eta μ mutazio-tasa *locus* eta belaunaldiko.

Locus bialelikoetan estimaturiko dibertsitate genetikoa

- Dibertsitate genetikoa analizatzeko modu erraza, ***locus* bakar** edo *loci* gutxi batzuen aldakortasunera mugatzea da.
- ***Loci* bialeliko**en, hots 2 alelo bakarrik dituztenak (A eta B aleloak adibidez) dibertsitatearen analisia biziki interesgarria da.
- *Loci* bialelikoetan, populazio batean azaltzen duten dibertsitatea neurtzeko modu bat populazio horretako lagin batetan behatzen diren AB gizabanako heterozigotoen portzentajea estimatzean datza.
- Aleloen maiztasunetatik abiatuta, esperotako heterozigositatea kalkulatu daiteke (**dibertsitate genikoa** deritzona). Ondoren, zenbait hautabide eskaintzen dira.

Populazioaren lagin batetik abiatuta dibertsitatearen estimazioa (θ)

Populazio batetako gizabanakoen lagin batetan, θ estimatzeko modu ezberdinak daude, adibidez:

1.- **Toki segregatzaileen** (edo polimorfiko) **kopurutik abiatuta, S.**

2.- **Batezbesteko ezberdintasun nukleotidikoen kopurutik abiatuta**

1.- θ-ren estimazioa **toki segregatzaileetatik abiatuta**

- Adibidez, demagun ondoko haplotipoez osotutako lagina:

  1 Haplotipoa : ACTGGC**T**AAGC**G**CAT**A**CTAG
  2 Haplotipoa: ACTGGCGAAGCCCATGCTAG
  3 Haplotipoa: AC**C**GG**T**GAAG**T**CCATGCT**T**G
  4 Haplotipoa: AC**C**GGCGAAGCCCATGCTAG

- Kasu honetan, S = 7 (haplotipoen artean aldaketaren bat gertatzen den toki kopurua).
- Watterson-ek (1975), populazio bat panmitikoa denean, orekan badago eta mutazioak neutroak badira, **S-ren esperotako balioa** (E(S)), ondokoa zela frogatu zuen:

$$E(S) = a_1\theta \qquad \text{non,} \quad a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \text{ eta } \theta = 4N_e\mu \text{ den}$$

- Beraz, bakandu ondoren, $\theta = S/a_1$ dela lortzen dugu. Honela, aurreko adibidean, S=7 eta n=4 izanik, a_1=1,8333 da. Ondorioz, θ -ren estima (θ_S deritzona) 7/ 1.8333 = 3,818 izango da.
- Hortaz, kasu honetan, θ_S = 3,818 da.

2.- θ-ren estimazioa bi sekuentzien arteko batezbesteko ezberdintasun nukleotidikoen kopuruaren bidez (k).

Bi sekuentzien arteko desberdintasun nukleotidikoak (k), honela definitzen dira:

$$k = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}}{n(n-1)}$$

k_{ij}-k sekuentzia guztiak binaka konparatu beharra (i vs j) suposatzen du eta zenbat toki nukleotidikotan ezberdintzatzen diren apuntatu, ondoren egindako konparaketa guztien batezbestekoa estimatzeko.

Gure kasuan, $k_{12} = 3$, $k_{13} = 7$, $k_{14} = 4$, $k_{23} = 4$, $k_{24} = 1$ eta $k_{34} = 3$. Beraz, batezbesteakoa, k, 2*(3+7+4+4+1+3)/(4*3) = 3.667 da.

k, θ_k deritzogun, θ-ren estima da.

K, toki nukleotidiko bakoitzerako esperotako heterozigositateen baturaren berdina da.

$$k = \sum_{i=1}^{S} h_i$$

Tajima-ren D testa (1989)

Tajimak proposatutakoa kontutan izanik, mutazioak selektiboki neutroak direnean, populazioa panmitikoa denean, orekan dagoenean eta ez badago errekonbinaziorik, k-ren esperotako balioa θ da. Baldintza hauetan, θ-ren estima biak gutxi gorabehera berdinak izan beharko lukete.

Aldiz, hautespena gertatzen denean (edo beste indar ebolutiboren batek eragiten badu), ez da aurrekoa beteko. Hortaz, Tajimak, θ -ren bi estimak konparatzen dituen test bat diseinatu zuen, zeina (Tajimaren) D testa deritzon eta hautespen indarren baten presentzia detektatzea baimentzen duen.

$$D = \frac{k - \frac{S}{a_1}}{\sqrt{e_1 S + e_2 S(S-1)}}$$ non,

$$e_1 = \frac{b_1 - 1/a_1}{a_1} \qquad e_2 = \frac{b_2 - (n+2)/(a_1 n) + a_2/a_1^2}{a_1^2 + a_2} \qquad a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \qquad a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2} \qquad b_1 = \frac{n+1}{3(n-1)} \qquad b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$$

Tajimaren D testaren interpretazioa

Tajimaren D estatistikoa, datuak (sekuentzia haplotipikoen bilduma), lagindu den populazioa eredu neutral baten eraginpean (mutazio-jito oreka) eboluzionatzen ari den ideiarekin bat datorren ebaluatzeko erabiltzen da.

- Populazioa mutazio-jito orekan badago *locus* horretarako (eboluzio neutrala, hautespenik ez), D-ren balioa ezin izango da 0tik modu estatistikoan desberdindu.

- Populazioa hautespen gainartzailearen eraginpean badago *locus* horretarako (heterozigotoaren abantaila), D›0 izango da.

- Populazioa hautespen purifikatzaile (mutazio berrien ezabapena) edo positiboaren (moldapena emendatzen duen mutazio berri baten alde) eraginpean badago *locus* horretarako, D‹0 izango da.

Orokorrean, D-ren balioa, detektatzen diren toki polimorfikoen kopuru eta maiztasunen menpe dago.

D-ren interpretazioa

• Hala ere, D interpretatzea konplexuagoa izan daiteke, besteak beste, aldaketa demografikoek ere D-rengan eragiten dutelako. Adibidez, populazio bat duela gutxi hasi bada hedatzen, D-ren balio negatiboa lortu dezakegu. Aldiz, populazio batek duela gutxi botila iduna jasan badu, D-k balio positiboak har ditzake.

• Gainera, D≠0 detektatzerakoan hutsegiteak gerta daitezke, laginaren tamaina behar bezain handia ez delako hautespenaren seinalea detektatu ahal izateko.

Dibertsitate genetikoa eta historia ebolutiboa

- Populazioen dibertsitate genetikoaren ereduak definitzen dituen beste aspektu garrantzitsu bat, populazioen **historia demografikoa** da.
- Gizakiotan, gure historia ebolutiboak zera adierazten du: *H. sapiens* Afrikan sortu zela duela gutxi gorabehera 200.000 urte, eta duela 100.000 urte inguru talde txiki bat Afrika utzi eta Ekialde Hurbilean kokatu zela, eta duela 40-50.000 urte inguru Europa eta Asia kolonizatu zituela, Europan bizi ziren neandertalak ordezkatuz.
- Beraz, espero duguna da afrikarren populazio tamaina handiago eta denbora luzeagoan *in situ* eboluzionatzen egon denez, populazio honen dibertsitatea eurasiarrena baino handiagoa izatea *locus* bererako.
- Printzipio honetatik edozein desbideratze, hautespenaren seinaletzat hartu daiteke.

Dibertsitate genetikoa analizatzeko tresna bioinformatikoak

3 programa jaitsi beharko ditugu (software askea)

Bioedit	http://www.mbio.ncsu.edu/bioedit/bioedit.html
DnaSP	http://www.ub.edu/dnasp/
Network	http://www.fluxus-engineering.com/sharenet.htm

Interneteko beste balibide batzuk

SPSmart	http://spsmart.cesga.es/
UCSC genome browser	https://genome-euro.ucsc.edu/

Berezko software-a

SPSmart2haps.exe (execute: c:\dir› SPSmart2haps.exe File-NAME-Root)
(requires a ref file: file-name-root_refseq_hg19.txt)

Aurretiaz, FASTA fitxategia zer den jakin behar dugu

FASTA fitxategia, indibiduo multzo bati dagozkion (gure populazioaren lagina) DNA sekuentzien (RNA edo proteinak ere izan daitezke) datuak dituen **testu bakuneko fitxategia** da. Datu hauek, **FASTA** deritzon formatu batetan adierazita daude.

Zer da testu bakuneko fitxategia?

"Bloc De Notas" (Windows) erabiltzerakoan sortzen den moduko fitxategia da. Testu bakuna, lodi, kolore,... gabe. Word erabilita ere sortu daiteke, baina "MS-Dos bezalako testu" moduan gorde behar dela zehaztu behar da.

Zer da FASTA formatua?

FASTA formatuan, sekuentzia bakoitza bi eremuz adierazita dago:

a) sekuentziaren identifikadorea
b) sekuentzia bera

Sekuentzien identifikadorea

Sekuentziari emango diogun hautazko izena da. Aurretik "›" ikurra darama eta bukatzeko lerro-jauzia (¶).

Sekuentzia bera

DNA sekuentzia haplotipikoa, IUB/IUPAC formatu estandarrean:

A adenina C zitosina G guanina T timina

N nukleotido ezezaguna - normalean base baten delezioa adierazten du

Letra xehe eta larriak onartzen dira.

Sekuentziak lerro bakarra zein zenbait lerro bete ditzake, karaktere kopuru finkoko segmentu jarraituetan banatzea erabakitzen bada ala ez kontutan izaink (normalean 80 karaktere lerroko). Sekuentzia guztiek formatu bera izan behar dute. Sekuentzia lerro-jauzi baten bidez bukatzen da (¶).

Sekuentzia desberdinak banatzeko, ez da lerro-hutsik onartzen.

FASTA formatuaren adibidea

›1 sekuentzia¶
ACTGACGATGACGATACAGTAGCGATGACGATGACGATAGGAG
ACAGCAGACGATATATAGACGATAGCAGTAGACGATACTGACG
ATGACGATACAGTAGCGATGAGATGACGATAGGAGACAGCAGA
CGCATATATAGACGATAGCAGTAGACGATACTGACT¶
›secuencia2¶
ACTGACGATGACGATACAGTAGCGATGACGATGACGATAGGAG
ACAGCAGACGATATATAGACGATAGCAGTAGAGGATACTGACG
ATGACGATACAGTAGCGATGAGATGACGATAGGAGACAGCAGA
CGCATATGTAGACGATAGCAGTAGACGATACTGACT¶
…

Edo segmentutan:

›1sek¶
ACTAGCATGACTGTGACGATGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGATGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGATGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
›2sek¶
ACTAGCATGACTGTGACGATGACGATGACGATGAT¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGTTGACGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGATGACGATAGCAG¶
ACTAGCATGACTGTGACGATGTCGATGACGATGAC¶
GCTCGGCTCGACAGFATAGCAGTTGACGATAGCAG¶
…

OHARRA

¶ sinboloak lerro-jauzia ("return" tekla) adierazten du; ez da idatzi behar. Lerro aldaketa adierazteko bistaratu dugu hemen. Fitxategia word-ekin irekiko bagenu eta "parrafo" atalean bistara dadila zehaztuko bagenu dagokion ikonoan klikatuz, agerikoa egingo litzateke.

FASTA fitxategiak .fas luzapena daramate ("fitxategia.fas" bezala, adibidez).

Jaitsiera eskema eta datuen formatua (1 eta 2 tutorialetan zehaztutako prozedura)

Tajimaren D testa bidezko hautespen analisiak egiteko (DnaSP programarekin) behar ditugun datuak modu egokian jaitsi eta formateatuko ditugu.

1.- Horretarako, interesatzen zaigun eskualde genomikoa mugatu beharko dugu, kasu honetan *MC1R* genea izango da. Bere koordenatu genomikoak (hg19) bilatu behar ditugu **UCSC genome Browser** erabiliz eta CDS-ari (UTR exoi gabe) dagokion sekuentzia genomikoa jaitsiko dugu. Testu bakuneko fitxategi bezala gordeko dugu (MC1R_refseq_hg19.txt). Aztertu.

2.- **SPSmart**–n web orrialdea ireki eta aurreko koordenatuak erabiliz, hasi datuen jaitsiera. Jaitsi afrikar populazioen multzoari dagozkion datuak (AFR) eta Europako populazio bakoitzari dagozkionak (bakoitza bere aldetik, GBR, FIN, CEU, IBS eta TS). **(3 tutorialean zehaztutako prozedura)**

3.- Egikaritu **SPSmart2haps.exe**, jaitsi den populazio bakoitzarekin (behin AFR-rekin , beste behin CEU-rekin, beste behin GBR etab...). Irteera fitxategietatik, bakoitzetik fitxategi.fas erabiliko dugu.

OHARRA: europar populazioetatik sortu bi fitxategi, baten Europaren Iparraldeko populazioak sartuko ditugu (NEU deritzona) eta bestean Europaren Hegoaldekoak sartuko ditugu (SEU deritzona). Horretarako, Bioedit programako "Merge Files" ezaugarria erabiliko dugu. Hiru fitxategi .fas sortuko ditugu: MC1R_AFR.fas, MC1R_NEU.fas y MC1R_SEU.fas

4.- Fitxategi hauetariko bakoitza ireki DnsSP-rekin eta kalkulatu **Tajimaren D balioa** (toki segregatzaileen kopurua erabiliz). Kalkulatu p-balioa simulazioen bidez ("coalescente" erabiliz). **(4 tutorialean zehaztutako prozedura)**

Bibliografia

- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*.1989; 123:585-595.

- Watterson GA. Number of segregating sites in genetic models without recombination. *Theor Popul Biol* (1975); 7: 256-276.

Gomendaturiko irakurketak

- Hartl DL & Clark AG. (2007) Principles of population genetics. 4th Ed. Sinauer.

- Joblin M, Hollox E, Hurles M, Kivisild T & Tyler-Smith C. (2014) Human Evolutionary Genetics. 2nd Edition, Garland Science. Pierce B.A. (2008) Genetics. A conceptual approach. 3rd Edition. Freeman and Co.