"सांख्यिकीय आनुवंशिकी और कृषि में इसके अनुप्रयोग" विषय पर ऑनलाईन हिंदी कार्यशाला

(मार्च 18 - 20, **2021**)

Hindi Online Workshop on

"Statistical Genetics and its Application in Agriculture"

(March 18 – 20, 2021)

समरेन्द्र दास **Samarendra Das**

उपेन्द्र प्रधान **Upendra Pradhan**

सन्दर्भ पुस्तिका

Reference Manual

सांख्यिकीय आनुवंशिकी प्रभाग

Division of Statistical Genetics

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, नई दिल्ली–**110012**

I.C.A.R.-Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi - 110012

सांख्यिकीय आनुवंशिकी और कृषि में इसके अनुप्रयोग

व्याख्यान सारणी

आर सॉफ्टवयर (R-software)

डॉ. समरेन्द्र दास
भा.कृ.अ.प.-भा.कृ.सां.अनु. संस्थान, नई दिल्ली-12

R आँकड़ों और ग्राफिक्स के लिए एक उच्च-स्तरीय कंप्यूटर भाषा और वातावरण है। यह विभिन्न प्रकार के सरल और उन्नत सांख्यिकीय तरीके निष्पादित करता है और उच्च गुणवत्ता वाले ग्राफिक्स का उत्पादन करता है। इसके अलावा, आर एक कंप्यूटर भाषा है, इसलिए, हम नए कार्यों को लिख सकते हैं जो आर के उपयोग का विस्तार करते हैं। शुरुआत में रॉस इहाका और रॉबर्ट जेंटलमैन द्वारा सांख्यिकी विभाग, ऑकलैंड विश्वविद्यालय, ऑकलैंड, न्यूजीलैंड (इसलिए नाम) में लिखा गया था।) का है। R एक कमांड संचालित सांख्यिकीय पैकेज है, जिसे 17 प्रोग्रामर के "R Core Team" सहित कई योगदानकर्ताओं द्वारा बनाए रखा गया है, जो R स्रोत कोड (R Core Team, 2012) को संशोधित करने के लिए जिम्मेदार हैं।

पहली नजर में, यह उपयोग करने के लिए इसे कठिन बना सकता है। हालाँकि, इस कंप्यूटर प्रोग्राम का उपयोग करके आँकड़े सीखने के कई कारण हैं। दो सबसे महत्वपूर्ण हैं:

a) आर मुक्त है; आप इसे http://www.r-project.org से डाउनलोड कर सकते हैं और इसे अपने पसंद के किसी भी प्रकार के कंप्यूटर पर स्थापित कर सकते हैं।

b) आर आपको उन सभी सांख्यिकीय परीक्षणों को करने की अनुमति देता है जिनकी आपको आवश्यकता है, सरल से उच्च उन्नत वाले तक। इसका मतलब है कि आपको हमेशा अपने डेटा पर सही विश्लेषण करने में सक्षम होना चाहिए।

इसके अलावा, आर में उत्कृष्ट ग्राफिक्स और प्रोग्रामिंग क्षमताएं हैं, इसलिए इसका उपयोग शिक्षण और सीखने में सहायता के रूप में किया जा सकता है। R की ताकत यह है कि सांख्यिकीय विश्लेषणों के साथ-साथ अच्छी तरह से डिज़ाइन किए गए प्रकाशन-गुणवत्ता वाले ग्राफिक्स का उत्पादन किया जा सकता है। आर सभी ऑपरेटिंग सिस्टम (लिनक्स, मैक और विंडोज) पर चलता है।

R डाउनलोड और पर्यावरण

R, CRAN मिरर साइटों (CRAN: व्यापक R संग्रह नेटवर्क) के नेटवर्क से आसानी से उपलब्ध है। R डाउनलोड करने और इंस्टॉल करने के लिए www.r-project.org पर जाएं और पास में एक CRAN मिरर चुनें। R एक कंसोल के माध्यम से संचालित कोड काम करता है, न कि उन मेनू के साथ जिनका उपयोग आप अन्य सॉफ़्टवेयर से कर सकते हैं। आर-कंसोल सिर्फ एक कैलकुलेटर है। अपने विश्लेषण के चरणों का दस्तावेजीकरण करने के लिए, आप अपने आर कोड को एक टेक्स्ट एडिटर में लिखेंगे (कोड के छोटे बिट्स को छोड़कर जिन्हें आपको सहेजने की आवश्यकता नहीं है)। पाठ संपादक से, आप कॉपी या भेज सकते हैं (यदि आपका संपादक आर के साथ बातचीत करता है) फ़ंक्शन कॉल को निष्पादित करने के लिए आर कंसोल को कोड। आप आर द्वारा उत्पादित परिणामों को पाठ फ़ाइलों में सहेज सकते हैं या विभिन्न प्रारूपों में ग्राफिक्स का उत्पादन कर सकते हैं। जब आप अपना R सत्र बंद करते हैं, तो R-कंसोल स्वयं सामान्य रूप से सहेजा नहीं जाता है। हालांकि, किसी भी समय अपने विश्लेषण को फिर से संगठित करने में सक्षम होने के लिए, आपको अपने आर कोड वाले टेक्स्ट फ़ाइल (फाइलों) को सहेजना चाहिए। यद्यपि आप आर कोड को लिखने और सहेजने के लिए किसी भी टेक्स्ट एडिटर का उपयोग कर सकते हैं (जैसे नोटपैड), यह एक टेक्स्ट एडिटर स्थापित करने की सिफारिश की गई है जो आर भाषा को पहचानता है, जैसे कि टिन-आर (http://www.sciviews.org/Tinn- R), RStudio (www.rstudio.org), या Emacs।

अपने कंप्यूटर पर आर स्थापित करने के बाद, यदि स्थापित नहीं है, तो http://www.r-project.org से मुफ्त में नवीनतम संस्करण डाउनलोड करें और आधार प्रणाली स्थापित करें। आपको अभी तक कोई अतिरिक्त पैकेज स्थापित करने की

आवश्यकता नहीं है। एक बार जब आप इसे स्थापित कर लेते हैं, तो इसे शुरू करें और आपको कुछ इस तरह प्रस्तुत करना चाहिए:

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

R में एक संपूर्ण सहायता "help.start ()" टाइप करके देखी जा सकती है। R में खोज इंजन जिसके बाद "कीवर्ड बाय टॉपिक" की सूची उपलब्ध है और इसे देखा जा सकता है।

आर की मूल बातें

इनपुट

यहां हम पता लगाते हैं कि आर सत्र में डेटा सेट को कैसे परिभाषित किया जाए। केवल दो आदेशों का पता लगाया जाता है। पहला डेटा के सरल असाइनमेंट के लिए है, और दूसरा डेटा फ़ाइल में पढ़ने के लिए है। आर सत्र में डेटा पढ़ने के कई तरीके हैं, लेकिन हम इसे सरल रखने के लिए सिर्फ दो पर ध्यान केंद्रित करते हैं।

संख्याओं की सूची को संग्रहीत करने का सबसे सीधा आगे तरीका सी कमांड का उपयोग करके असाइनमेंट के माध्यम से है। " सी कमांड के साथ एक सूची निर्दिष्ट की गई है, और असाइनमेंट "<-" प्रतीकों के साथ निर्दिष्ट किया गया है। संख्याओं की सूची का वर्णन करने के लिए उपयोग किया जाने वाला एक और शब्द इसे "वेक्टर" कहना है। सी कमांड के भीतर संख्याओं को कॉमा द्वारा अलग किया जाता है।

उदाहरण के लिए, हम "a" नामक एक नया चर बना सकते हैं जिसमें 3, 5, 7 और 9 नंबर होंगे:

```
> a <- c(3,5,7,9)
```

जब आप इस कमांड को दर्ज करते हैं तो आपको नई कमांड लाइन को छोड़कर कोई आउटपुट नहीं देखना चाहिए। कमांड संख्या की एक सूची बनाता है जिसे "ए" कहा जाता है। यह देखने के लिए कि "a" में कौन सी संख्याएँ शामिल हैं, बस "a" टाइप करें और एंटर की दबाएं और परिणाम होगा:

```
> a
[1] 3 5 7 9
```

यदि आप संख्याओं में से किसी एक के साथ काम करना चाहते हैं, तो आप चर का उपयोग करके इसे प्राप्त कर सकते हैं और फिर वर्ग कोष्ठक जो यह दर्शाता है कि कौन सी संख्या:

```
> a[2]
[1] 5
```

1.2। डेटा फ़ाइल पढ़ना

दुर्भाग्य से, यह केवल कुछ डेटा बिंदुओं के लिए दुर्लभ है, जिन्हें आपको प्रॉम्प्ट पर टाइप करने में कोई आपत्ति नहीं है। जटिल संबंधों (जैसे जीनोमिक डेटा) के साथ बहुत अधिक डेटा बिंदु होना बहुत आम है। यहां हम यह जांचेंगे कि रीड.टेबल और अन्य फ़ंक्शन का उपयोग करके किसी फ़ाइल से डेटा सेट कैसे पढ़ें, लेकिन पहले डेटा फ़ाइल कैसे बनाएं।

data.frame फ़ंक्शन डेटा फ़्रेम बनाता है, चर के कसकर युग्मित संग्रह जो आर के अधिकांश मॉडलिंग सॉफ़्टवेयर द्वारा मूलभूत डेटा संरचना के रूप में उपयोग किए जाने वाले मैट्रिसेस और सूचियों के कई गुणों को साझा करता है।

```
data.frame(..., row.names = NULL, check.rows = FALSE,
        check.names = TRUE,
        stringsAsFactors = default.stringsAsFactors())
default.stringsAsFactors()
```

read.table तालिका प्रारूप में एक फ़ाइल पढ़ता है और फ़ाइल में फ़ील्ड के लिए लाइनों और चर के अनुरूप मामलों के साथ, इससे एक डेटा फ़्रेम बनाता है।

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
        dec = ".", row.names, col.names,
        as.is = !stringsAsFactors,
        na.strings = "NA", colClasses = NA, nrows = -1,
        skip = 0, check.names = TRUE, fill = !blank.lines.skip,
        strip.white = FALSE, blank.lines.skip = TRUE,
        comment.char = "#",
        allowEscapes = FALSE, flush = FALSE,
        stringsAsFactors = default.stringsAsFactors(),
        encoding = "unknown")
```

लिखना: डेटा (आमतौर पर एक मैट्रिक्स) x फ़ाइल फ़ाइल के लिए लिखा जाता है। यदि x एक द्वि-आयामी मैट्रिक्स है, तो इसे आंतरिक प्रतिनिधित्व के रूप में फ़ाइल में कॉलम प्राप्त करने के लिए इसे स्थानांतरित करने की आवश्यकता हो सकती है।

```
write (x, file = "data", ncolumns = if(is.character(x)) 1 else 5, append = FALSE, sep = " ")
```

X	डेटा बाहर लिखा जाना है
File	एक कनेक्शन, या एक चरित्र स्ट्रिंग को लिखने के लिए फ़ाइल का नामकरण। यदि "", मानक आउटपुट कनेक्शन पर प्रिंट करें।
ncolumns	डेटा लिखने के लिए कॉलम की संख्या।
append	यदि TRUE डेटा x को कनेक्शन से जोड़ा जाता है।
Sep	स्तंभों को अलग करने के लिए प्रयुक्त एक स्ट्रिंग। Sep = "\ t" का उपयोग टैब सीमांकित आउटपुट देता है; डिफ़ॉल्ट "" है।

लिखने में सक्षम। अपने आवश्यक तर्क एक्स को प्रिंट करता है (एक फ़ाइल या कनेक्शन के लिए यह एक डेटा फ्रेम में बदलने के बाद अगर यह एक और न ही मैट्रिक्स है)।

```
write.table(x, file = "", append = FALSE, quote = TRUE, sep = " ",
        eol = "\n", na = "NA", dec = ".", row.names = TRUE,
        col.names = TRUE, qmethod = c("escape", "double"))
```

2. बुनियादी डेटा प्रकार

2.1। चर प्रकार

२.१.१। नंबर

वास्तविक संख्याओं के साथ काम करने का तरीका पहले से ही पहले अध्याय में शामिल किया गया है और यहां संक्षेप में चर्चा की गई है। किसी संख्या को संग्रहीत करने का सबसे मूल तरीका एक संख्या का असाइनमेंट बनाना है:

```
a <- 3
```

"<-" आर को प्रतीक के दाईं ओर संख्या लेने और एक चर में संग्रहीत करने के लिए कहता है जिसका नाम बाईं ओर दिया गया है। आप "=" प्रतीक का भी उपयोग कर सकते हैं। जब आप एक असाइनमेंट बनाते हैं तो आर किसी भी जानकारी को प्रिंट नहीं करता है।

यदि आप यह देखना चाहते हैं कि किसी चर का मान किसी रेखा पर चर का नाम किस प्रकार है और एंटर की दबाएं:

```
> a
[1] 3
```

यह आपको सभी प्रकार के बुनियादी कार्यों को करने और संख्याओं को बचाने की अनुमति देता है:

```
> b <- sqrt(a*a+3)
> b
[1] 3.464102
```

२.१.२। स्ट्रिंग्स

आप केवल स्टोरिंग नंबर तक सीमित नहीं हैं। आप स्ट्रिंग्स को स्टोर भी कर सकते हैं। उद्धरण का उपयोग करके एक स्ट्रिंग निर्दिष्ट की जाती है। दोनों सिंगल और डबल कोट्स काम करेंगे:

```
> a <- "hello"
> a
[1] "hello"
> b <- c("hello","there")
> b
[1] "hello" "there"
> b[1]
[1] "hello"
> typeof(a)
[1] "character"
> a = character(20)
> a "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
```

२.१.३। कारकों

एक और महत्वपूर्ण तरीका, आर डेटा को एक कारक के रूप में संग्रहीत कर सकता है। अक्सर एक प्रयोग में कुछ व्याख्यात्मक चर के विभिन्न स्तरों के लिए परीक्षण शामिल होते हैं। उदाहरण के लिए, जब एक पेड़ की वृद्धि दर पर कार्बन डाइऑक्साइड के प्रभाव को देखते हुए आप यह देखने की कोशिश कर सकते हैं कि कार्बन डाइऑक्साइड के अलग-अलग पूर्व निर्धारित सांद्रता के संपर्क में आने पर विभिन्न पेड़ कैसे बढ़ते हैं। विभिन्न स्तरों को कारक भी कहा जाता है।

नीचे दिए गए उदाहरण (उदाहरण के लिए पेड़ डेटा फ़ाइल) के लिए, फ़ाइल में कई चर कारक हैं:

```
> summary(tree$CHBR)

A1  A2  A3  A4  A5  A6  A7  B1  B2  B3  B4  B5  B6  B7  C1  C2  C3  C4  C5  C6
3   1   1   3   1   3   1   1   3   3   3   3   3   3   1   3   2   3   1   1
C7 CL6 CL7  D1  D2  D3  D4  D5  D6  D7
1   1   1   1   1   3   1   1   1   4
```

इस डेटा सेट में कई स्तंभ कारक हैं, लेकिन शोधकर्ताओं ने विभिन्न स्तरों को इंगित करने के लिए संख्याओं का उपयोग किया। उदाहरण के लिए, "A1" लेबल वाला पहला कॉलम एक कारक है। प्रत्येक पेड़ एक ऐसे वातावरण में उगाया गया था जिसमें कार्बन डाइऑक्साइड के चार अलग-अलग संभावित स्तरों में से एक था। शोधकर्ताओं ने काफी समझदारी से इन चार वातावरणों को 1, 2, 3 और 4 के रूप में लेबल किया। दुर्भाग्य से, आर यह निर्धारित नहीं कर सकता है कि ये कारक हैं और उन्हें यह मान लेना चाहिए कि वे नियमित संख्या हैं।

3. बुनियादी संचालन और संख्यात्मक विवरण

हम कुछ बुनियादी ऑपरेशनों को देखते हैं जिन्हें आप संख्याओं की सूची पर कर सकते हैं। यह माना जाता है कि आप डेटा दर्ज करना जानते हैं या डेटा फ़ाइलों को पढ़ना चाहते हैं जो उपरोक्त अनुभाग में शामिल हैं और आपको मूल डेटा प्रकारों के बारे में पता है।

3.1। बुनियादी संचालन

एक बार जब आपके पास एक वेक्टर (या संख्याओं की एक सूची) स्मृति में सबसे बुनियादी संचालन उपलब्ध हैं। अधिकांश बुनियादी ऑपरेशन एक पूरे वेक्टर पर कार्य करेंगे और एक ही आदेश के साथ बड़ी संख्या में गणना करने के लिए जल्दी से उपयोग किए जा सकते हैं। ध्यान देने वाली एक बात है, यदि आप एक से अधिक वेक्टर पर एक ऑपरेशन करते हैं तो अक्सर यह आवश्यक होता है कि वैक्टर सभी में समान संख्या में प्रविष्टियाँ हों।

मूल उदाहरण

R कोड लाइन को लाइन से चलाता है। यही है, आप इसे एक बात बताते हैं, और यह इसे तुरंत करता है। (कभी-कभी यदि कोड की हमारी एक "लाइन" सुपर लंबी होती है, तो यह वास्तव में एक पृष्ठ पर कई लाइनों के रूप में लिखा जाएगा, लेकिन आर इसे कोड के एक सुपर-लॉन्ग वाक्य के रूप में मानता है)।

संख्याओं के साथ, हम कैलकुलेटर की तरह आर का उपयोग कर सकते हैं। जब हम 3 + 7 टाइप करते हैं और एंटर करते हैं, तो कंसोल विंडो में जो दिखाई देता है, उसका एक उदाहरण निम्नलिखित है।

```
> 3+7
[1] 10
```

Basic arithmetic operators	code	Results
+	3+7	3+7=10
-	3-7	3-7=-4
*	3*7	3X7=21
/	3/7	3/7=0.4286
sqrt	sqrt(3)	√3=1.732045
log	log(2)	Natural Logarithm
exp	exp(log(2))	2
sin, cos, tan	sin(a), cos(a), tan(a)	sin(a), cos(a), tan(a)
$\sin^{-1}$, $\cos^{-1}$, $\tan^{-1}$	asin(a),a cos(a),atan(a)	asin(a),a cos(a),atan(a)

हम नामों का उपयोग करके वस्तुओं को संग्रहीत भी कर सकते हैं। हम इस वर्ग में नामांकित डेटा फ़्रेम के साथ सबसे अधिक बार देखते हैं। (उर्फ डेटा सेट)। हम तालिकाओं, फ़ंक्शन आउटपुट या एकल मान भी संग्रहीत करेंगे। एक सरल उदाहरण निम्नलिखित कोड है:

se <- sqrt(.75*.25/200)

उदाहरण के लिए: मैं अपने कार्यक्षेत्र में "से" के रूप में 200 टिप्पणियों के साथ .75 के नमूने अनुपात के लिए मानक त्रुटि को संग्रहीत करना चाहता हूं। यह सुविधाजनक है अगर मैं इसे समीकरणों में बार-बार उपयोग करने जा रहा हूं। आप देखेंगे कि यदि आप कोड की इस लाइन को चलाते हैं, तो आपके कंसोल में कोई आउटपुट दिखाई नहीं देता है। लेकिन आपके कार्यक्षेत्र में एक नया "मूल्य" प्रकट होता है, जिसे se कहा जाता है। आप "<-" के बजाय मान निर्दिष्ट करने के लिए "=" का उपयोग कर सकते हैं। पाठ्यपुस्तक "=" का उपयोग करती है, लेकिन कई एक सम्मेलन के रूप में तीर का उपयोग करना पसंद करते हैं; जैसा कि आप अधिक कोड लिखते हैं, आप अपनी शैली विकसित करेंगे।

Note that R is case sensitive. The object se is not the same as SE.

इसके अलावा, आर सॉफ्टवेयर अनुसंधान के सभी क्षेत्रों से प्राप्त अधिकांश प्रयोगात्मक डेटा का विश्लेषण करता है। विवरणात्मक सांख्यिकी, प्रतिगमन, सहसंबंध, रैखिक मॉडल, विचरण का विश्लेषण, पूरी तरह से यादृच्छिक डिजाइन, यादृच्छिक पूर्ण ब्लॉक डिजाइन, लैटिन वर्ग डिजाइन, प्रमुख घटक विश्लेषण, क्लस्टर विश्लेषण, आदि जैसे सांख्यिकीय विश्लेषण आर। विश्लेषण में उपलब्ध हैं जो उपरोक्त तकनीकों पर आधारित हैं। व्यावहारिक रूप से, विस्तार से वास्तविक जीवन के उदाहरणों से निपटा जाता है।

इन विश्लेषणों के बारे में मदद आसानी से help.start () को शेल प्रॉम्प्ट पर टाइप करके और संक्षेप में दी जा सकती है:

कक्षाएं: डेटा प्रकार

ओ एनए: गुम मान

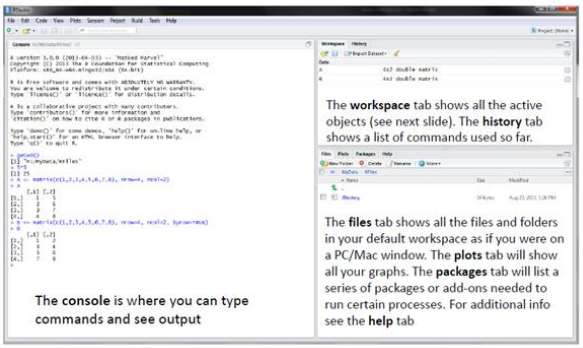
ओ श्रेणी: श्रेणीबद्ध डेटा
ओ चरित्र: चरित्र डेटा ("स्ट्रिंग") संचालन
o जटिल: जटिल संख्या
• डेटा: वातावरण, स्कोपिंग, पैकेज
• डेटासेट: डेटा द्वारा उपलब्ध डेटासेट ()
• सूची: सूचियाँ
• हेरफेर: डेटा हेरफेर
• पैकेज: पैकेज सारांश
• sysdata: बुनियादी प्रणाली चर
ग्राफिक्स
• aplot: मौजूदा प्लॉट / आंतरिक भूखंड में जोड़ें
• रंग: रंग, पट्टियाँ आदि
• डिवाइस: ग्राफिकल डिवाइस
• dplot: प्लॉटिंग से संबंधित संगणना
• गतिशील: गतिशील ग्राफिक्स
• hplot: उच्च-स्तरीय भूखंड
• ipl: प्लॉट के साथ बातचीत
MASS (पुस्तक) का उपयोग करता है
• वर्गीकरण: वर्गीकरण
• तंत्रिका: तंत्रिका नेटवर्क
• स्थानिक: स्थानिक सांख्यिकी
गणित
• एरीथ: बेसिक अंकगणित और छंटनी
• सरणी: मैट्रिसेस और एरेस
o बीजगणित: रैखिक बीजगणित
• रेखांकन: रेखांकन (ग्राफिक्स नहीं), यानी नोड्स

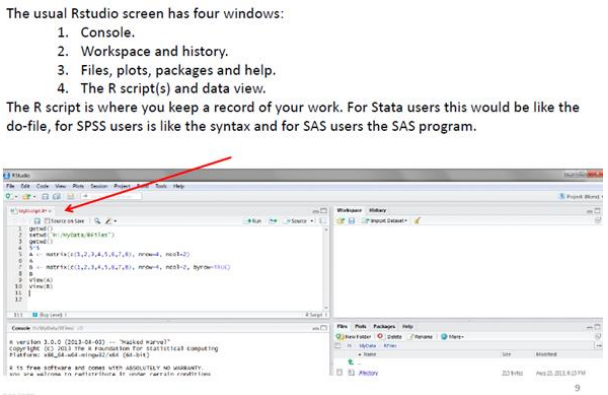
प्रोग्रामिंग, इनपुट / Ouput, और विविध
• IO: इनपुट / आउटपुट
ओ कनेक्शन: इनपुट / आउटपुट कनेक्शन
ओ डेटाबेस: डेटाबेस के लिए इंटरफेस
ओ फाइल: इनपुट / आउटपुट फाइलें
• डीबगिंग: डीबगिंग टूल
• प्रलेखन: प्रलेखन
• पर्यावरण: सत्र पर्यावरण
• त्रुटि: त्रुटि हैंडलिंग
• आंतरिक: आंतरिक ऑब्जेक्ट (एपीआई का हिस्सा नहीं)
• पुनरावृति: लूपिंग और पुनरावृत्ति

• तरीके: तरीके और सामान्य कार्य
• विविध: विविध
आंकड़े
• क्लस्टर: क्लस्टरिंग
डेटा सेट उत्पन्न करने के लिए कार्य
• डिजाइन: डिजाइन प्रयोगों
• वितरण: संभाव्यता वितरण और यादृच्छिक संख्या
• htest: सांख्यिकीय इंजेक्शन
• मॉडल: सांख्यिकीय मॉडल
ओ प्रतिगमन: प्रतिगमन
 नॉनलाइनर: गैर-रेखीय प्रतिगमन
• बहुभिन्नरूपी: बहुभिन्नरूपी तकनीक
• नॉनपैरेमेट्रिक: नॉनपैरेमेट्रिक सांख्यिकी
• मजबूत: मजबूत / प्रतिरोधी तकनीक
• चिकनी: वक्र (और सतह) चौरसाई
ओ loess: ढीली वस्तुओं
RStudio

RStudio सांख्यिकीय प्रोग्रामिंग सॉफ़्टवेयर R के लिए एक उपयोगकर्ता इंटरफ़ेस है। जबकि कुछ ऑपरेशन माउस से इंगित और क्लिक करके किए जा सकते हैं, प्रोग्राम कोड लिखना सीखना आवश्यक है। यह एक नई भाषा सीखने की तरह है - विशिष्ट वाक्यविन्यास, व्याकरण और शब्दावली है, और इसका उपयोग करने में समय लगेगा। आर स्टूडियो सीखना अंततः आर पर डेटा का विश्लेषण और कल्पना करते समय पूर्ण नियंत्रण, लचीलापन और रचनात्मकता देगा, लेकिन इस नई भाषा में प्रवाह में समय लगेगा।

एक बार जब आप R डाउनलोड कर लेते हैं, तो आप http://www.rstudio.com/ से RStudio स्थापित कर सकते हैं, "अभी डाउनलोड करें" पर क्लिक करके, और फिर "RStudio डेस्कटॉप डाउनलोड करें" पर क्लिक करें। अपने ऑपरेटिंग सिस्टम के लिए उपयुक्त संस्करण का चयन करें और डाउनलोड करें। जब आप RStudio खोलेंगे तो आपको निम्न स्क्रीन दिखाई देगी और चार विंडो होंगी:





चार खिड़कियों के रूप में वर्णित किया जा सकता है:
लिपि

स्क्रिप्ट R कमांड की एक सूची को संग्रहीत करने के लिए एक दस्तावेज़ है। जब आप पहली बार RStudio खोलते हैं तो यह विंडो प्रकट नहीं हो सकती है। एक नई स्क्रिप्ट बनाने के लिए, "फाइल -> नया -> आर स्क्रिप्ट" पर क्लिक करें।

कंसोल

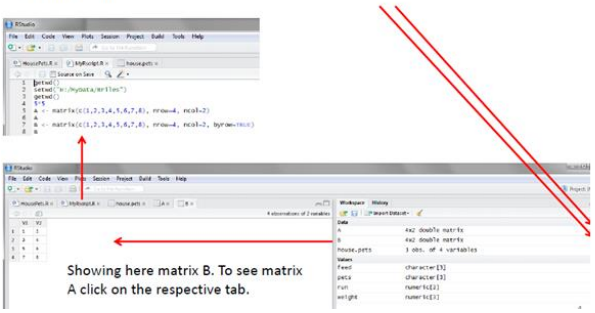
यहां आउटपुट दिखाई देता है। > संकेत (जिसे "प्रॉम्प्ट" भी कहा जाता है) का अर्थ है कि आर आज्ञाओं को स्वीकार करने के लिए तैयार है। आप कमांड को सीधे कंसोल में टाइप कर सकते हैं। हालाँकि, इसके बजाय स्क्रिप्ट विंडो में टाइप करना और वहाँ से कमांड चलाना एक अच्छी आदत है। कंसोल में कुछ भी नहीं बचाया जा सकता है। हालाँकि आप अपने आदेश को स्क्रिप्ट फ़ाइल में सहेज सकते हैं, और फिर बाद में अपने विश्लेषण को दोहरा सकते हैं। यदि आप किसी बड़ी परियोजना पर काम कर रहे हैं या आप बाद में वापस आने के लिए अपना कोड रखना चाहते हैं तो यह विशेष रूप से सहायक है।

कार्यस्थान

यह कार्यस्थान विंडो आपके पास वर्तमान में उपलब्ध वस्तुओं को सूचीबद्ध करती है। फ़ंक्शंस जो "बेस आर" या पैकेज का हिस्सा हैं, वे यहां दिखाई नहीं देंगे (उस प्रैक्टिकल को बनाने के लिए बस बहुत सारे हैं!) विशेष फ़ंक्शन जो आप खुद लिखते हैं या जो हैं।

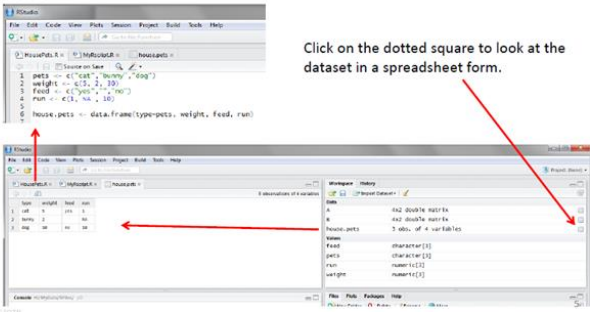
Workspace tab (1)

The workspace tab stores any object, value, function or anything you create during your R session. In the example below, if you click on the dotted squares you can see the data on a screen to the left.



Workspace tab (2)

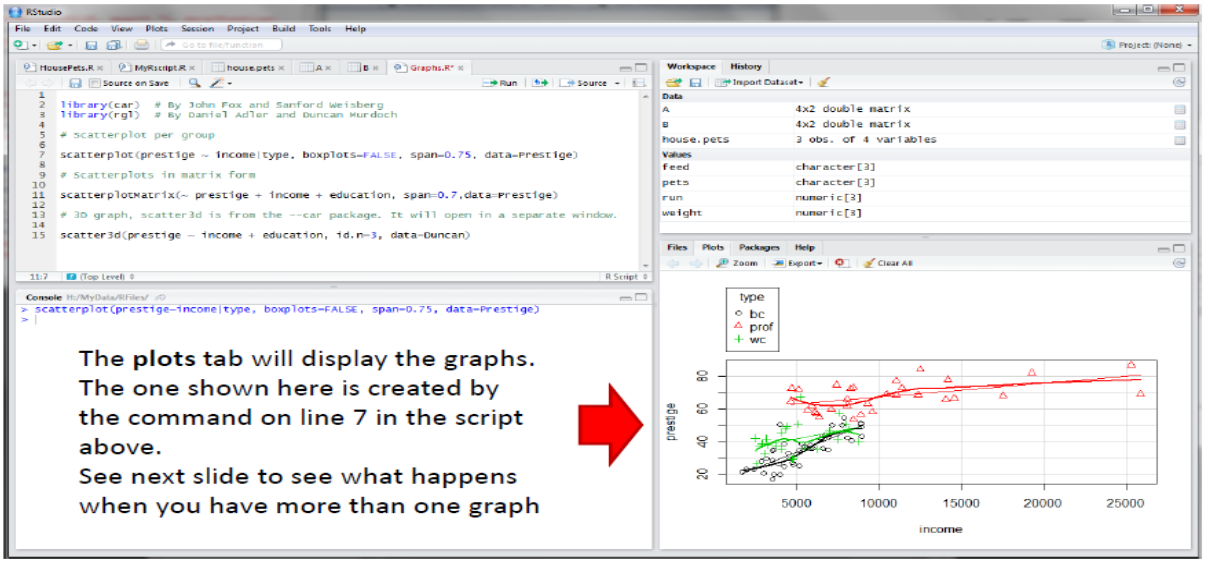
Here is another example on how the workspace looks like when more objects are added. Notice that the data frame `house.pets` is formed from different individual values or vectors.



प्लॉट / सहायता

अंतिम विंडो में कई टैब हैं, जिसमें एक खोज सुविधा के साथ एक सहायता टैब भी शामिल है। जब आप भूखंड बनाते हैं तो वे इस विंडो में दिखाई देंगे, जिसे आप बेहतर दृश्य प्राप्त करने के लिए आकार बदल सकते हैं। "फ़ाइलें" टैब आपको आपके द्वारा पहले लिखी गई आर लिपियों तक पहुंचने के एक तरीके के रूप में आपके कंप्यूटर पर फाइलें भी दिखाता है। सावधान रहें- इस विंडो में फ़ाइलों को हटाने से उन्हें आपके कंप्यूटर से हटा दिया जाता है। आर स्टूडियो में प्लॉट विंडो के रूप में कल्पना की जा सकती है:

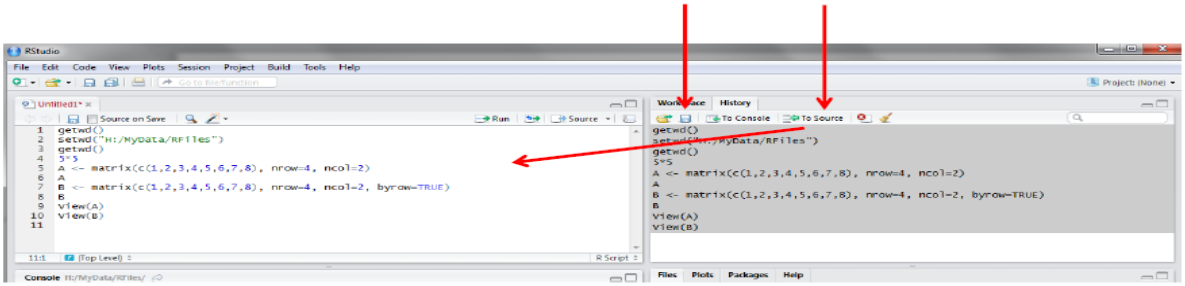
Plots tab (1)



इसके अलावा, इतिहास की खिड़की को अच्छी तरह से देखा जा सकता है:

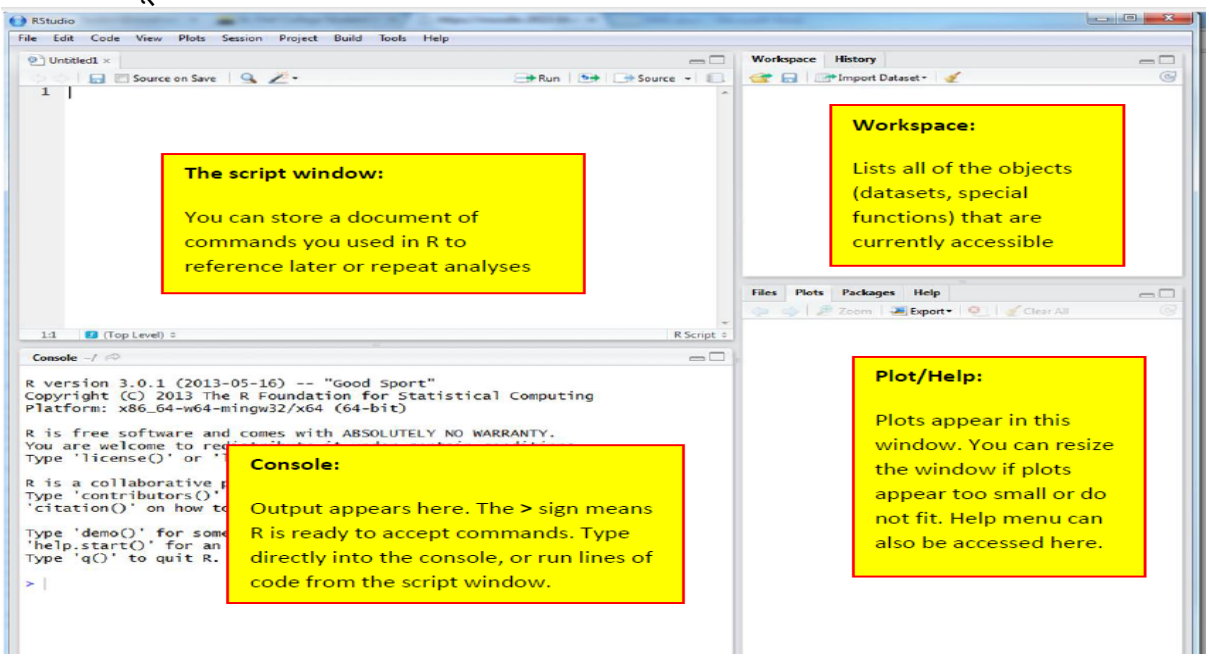
History tab

The history tab keeps a record of all previous commands. It helps when testing and running processes. Here you can either **save** the whole list or you can **select** the commands you want and send them to an R script to keep track of your work.
In this example, we select all and click on the "To Source" icon, a window on the left will open with the list of commands. Make sure to save the 'untitled1' file as an *.R script.



6

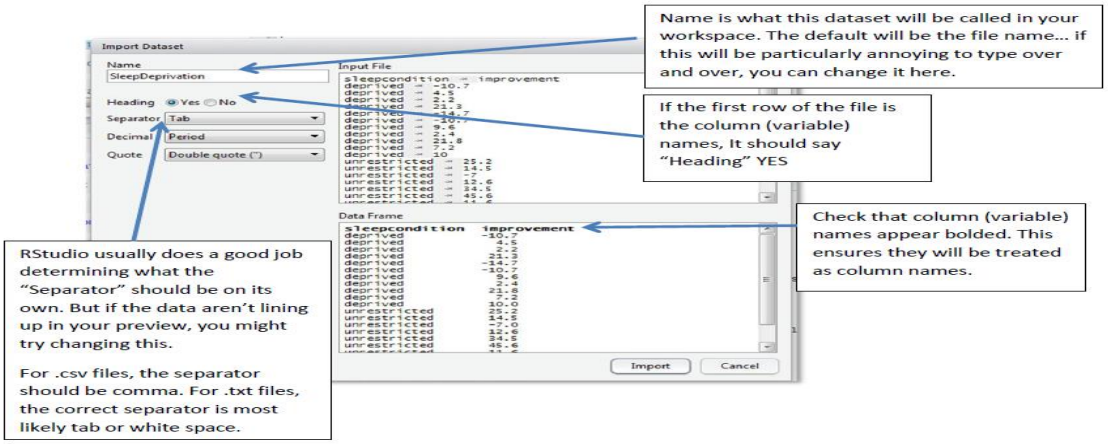
हालाँकि, R स्टूडियो की इन चार खिड़कियों को अच्छी तरह से देखा जा सकता है:



RStudio में कार्यस्थान और डेटासेट लोड हो रहे हैं

.RData फ़ाइल एक्सटेंशन का उपयोग करके कार्यस्थानों को सहेजा जाता है। एक कार्यक्षेत्र कई डेटासेट या आपके द्वारा लिखे गए कार्यों के एक सेट को स्टोर करने का एक सुविधाजनक तरीका है, खासकर जब कोड को चलाने के लिए डेटासेट का उत्पादन करने में लंबा समय लग सकता है। कार्यस्थान लोड करने के लिए, ऊपरी दाएँ RStudio विंडो में "कार्यस्थान" टैब के अंतर्गत फ़ोल्डर आइकन पर क्लिक करें। जहाँ भी आपने कार्यक्षेत्र को सहेजा है और उसे खोलें, पर नेविगेट करें। अब आपको कार्यक्षेत्र में वस्तुओं की एक सूची देखनी चाहिए । अपने कार्यक्षेत्र में डेटासेट लोड करने के लिए, आपको आयात डेटा बटन पर क्लिक करने और "फ़ाइल से" या "URL से" उपयुक्त के रूप में चयन करने की आवश्यकता है। आप csv या txt फ़ाइलों को लोड कर सकते हैं जिन्हें आपने "फ़ाइल से" अपने कंप्यूटर पर सहेजा है। जब डेटा ऑनलाइन पाठ फ़ाइलों के रूप में दिखाई देते हैं, तो आप उन्हें सीधे URL से लोड करने में सक्षम हो सकते हैं।

अब आपको केवल यह सुनिश्चित करने की आवश्यकता है कि डेटा फ़्रेम का पूर्वावलोकन सही है या नहीं और जैसा दिखाया गया है:



Once you import the dataset, a new data frame will appear in your workspace with whatever name was in the "Name" box.

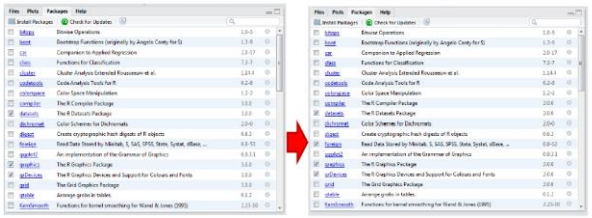
संकुल

जबकि कई उपयोगी फ़ंक्शंस "बेस आर" में शामिल हैं, उपयोगकर्ता और डेवलपर्स विशेष कार्यों और डेटासेट के साथ अपने स्वयं के ऐड-ऑन पैकेज बना और जमा कर सकते हैं। इन पैकेजों तक पहुँचने के लिए दो चरणों की आवश्यकता होती है: पैकेज को अपने कंप्यूटर पर स्थापित करना (केवल एक बार करने की आवश्यकता होती है) और अपने कार्यक्षेत्र में पुस्तकालय को लोड करने की आवश्यकता है (हर बार जब आप RStudio खोलते हैं तो ऐसा करने की आवश्यकता होती है)।

संकुल बिंदु द्वारा स्थापित किया जा सकता है और RStudio में क्लिक कर सकते हैं।

Packages tab

The package tab shows the list of add-ons included in the installation of RStudio. If checked, the package is loaded into R, if not, any command related to that package won't work, you will need select it. You can also install other add-ons by clicking on the 'Install Packages' icon. Another way to activate a package is by typing, for example, `library(foreign)`. This will automatically check the `--foreign` package (it helps bring data from proprietary formats like Stata, SAS or SPSS).



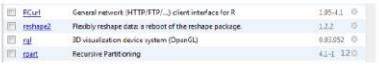
Installing a package

Before

Also can be installed by typing install.packages"pkg_name" on R console

Click on "`Install Packages`", write the name in the pop-up window and click on "`Install`".

After

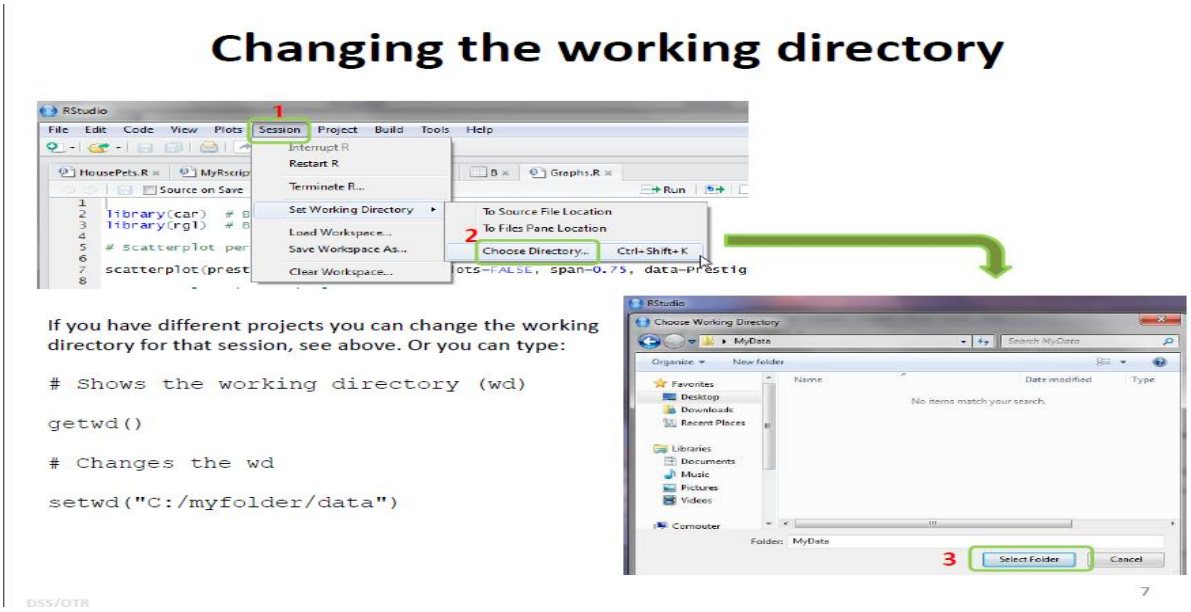


कार्यशील निर्देशिका बदलना

R को हमेशा कंप्यूटर पर एक निर्देशिका में इंगित किया जाता है। यह आसानी से पता लगाया जा सकता है कि कौन सी डायरेक्टरी गेटवे (वर्किंग डायरेक्टरी प्राप्त करें) फ़ंक्शन को चलाकर; इस फ़ंक्शन का कोई तर्क नहीं है। कार्यशील निर्देशिका को बदलने के लिए, सेटवार्ड का उपयोग करें और वांछित फ़ोल्डर में पथ निर्दिष्ट करें। dir - एक कार्यशील निर्देशिका निर्दिष्ट करें। इसके अलावा, getwd R प्रक्रिया की वर्तमान कार्यशील निर्देशिका का प्रतिनिधित्व करते हुए एक निरपेक्ष फ़ाइलपथ लौटाता है; setwd (dir) का उपयोग कार्य निर्देशिका को dir में सेट करने के लिए किया जाता है।

Usage

```
getwd()
setwd(dir)
```



स्क्रिप्ट विंडो में कोड लिखना

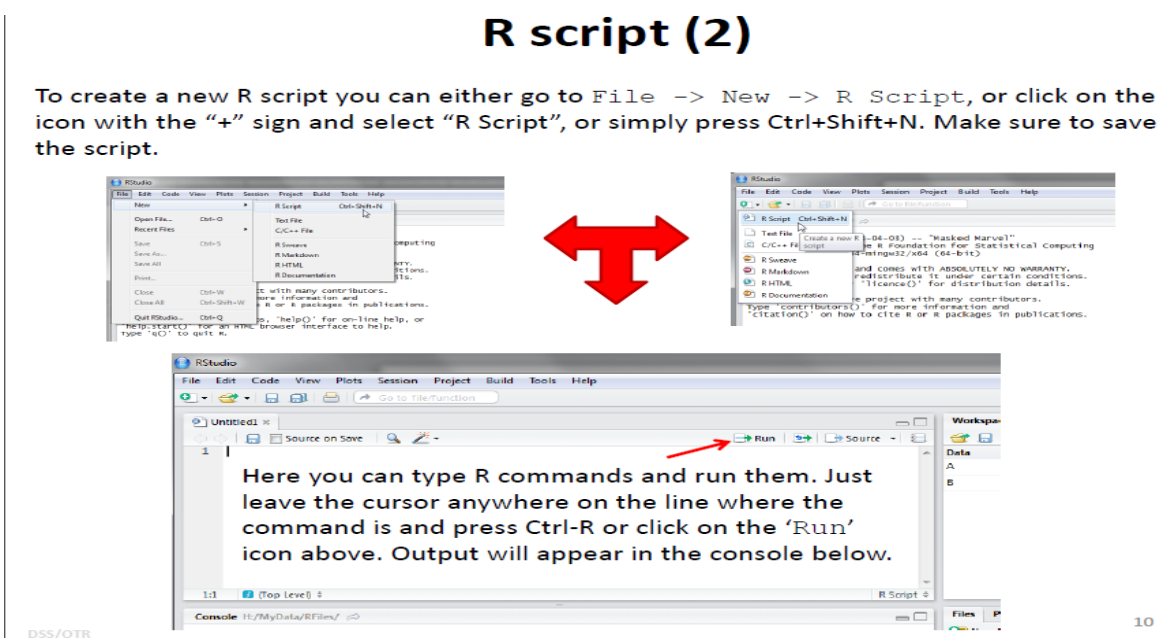
यह वास्तव में सीधे सांत्वना में सब कुछ टाइप करने के लिए आकर्षक हो सकता है- और यदि आप केवल एक या दो लाइनों का विश्लेषण कर रहे हैं जिसे आप कभी नहीं दोहराएंगे, तो यह ठीक हो सकता है। हालांकि, होमवर्क और प्रोजेक्ट्स करते समय आपके द्वारा चलाए गए कोड की एक प्रति होना आवश्यक होगा। मैं अक्सर आपके साथ R स्क्रिप्ट साझा करूंगा जिसमें उदाहरण शामिल हैं। इन्हें रखना एक अच्छा विचार है, और यहां तक कि अपनी टिप्पणी और नोट्स भी जोड़ें क्योंकि हम उन्हें कक्षा में उपयोग करते हैं।

स्क्रिप्ट विंडो में कोड लिखने के लाभ:

1. ऑटो-रंग और कोष्ठक हाइलाइटिंग त्रुटियों को खोजने में आसान बनाते हैं।
2. आप अपने कोड को सहेज सकते हैं और बाद में संदर्भ के लिए खुद को नोट्स लिख सकते हैं।
3. प्रोजेक्ट्स पर काम करते समय, या प्रश्न होने पर मेरे साथ साझा करने के लिए सहपाठियों के साथ अपना कोड साझा करना आसान बनाता है।
4. अपने विश्लेषण को दोहराने योग्य बनाता है, संपादित करने और कॉपी करने में आसान।

लिपियों में फ़ाइल एक्सटेंशन ".R" होता है, यदि आपने अपने कंप्यूटर पर R स्थापित नहीं किया है, तो आप एक। संपादक जैसे नोटपैड (हालांकि तब आप केवल सादे पाठ, रंगों को नहीं देख सकते हैं) का उपयोग करके .R फाइलें देख सकते हैं। सादा पाठ फ़ाइलें (.txt) भी स्क्रिप्ट फ़ाइलों के रूप में RStudio में खोली जा सकती हैं।

स्क्रिप्ट फ़ाइलों के बारे में महान चीजों में से एक टिप्पणी शामिल करने की क्षमता है। ये R कमांड के साथ डाले गए नोट हैं जो R में नहीं चलेंगे।



आम त्रुटि संदेश आर में

यदि आपको लाल आउटपुट मिलता है, तो आपने एक त्रुटि का अनुभव किया है। यहां कुछ सबसे आम त्रुटि संदेश दिए गए हैं जिनका आप सामना करेंगे।

Error: Object '...' not found

इसका मतलब है कि संदर्भित वस्तु आपके कार्यक्षेत्र में नहीं है। यह हो सकता है क्योंकि:

1. आप डेटा लोड करना या पैकेज स्थापित करना भूल गए।

2. आपने टाइप-ओ या कैपिटलाइज़ेशन त्रुटि की है।

3. आप उद्धरण चिह्नों को भूल गए होंगे, उदा। परिकल्पना परीक्षणों के लिए फ़ंक्शन इनपुट के रूप में "अधिक"। यह भी जाँचें कि तार्किक जैसे TRUE / FALSE सभी कैप में हैं।

4. आप पहले कोड की एक पंक्ति चलाना भूल गए थे, जिस ऑब्जेक्ट का आप उल्लेख कर रहे हैं। (सुनिश्चित करने के लिए अपने कंसोल से स्क्रॉल करें)।

5. आप एक विशिष्ट डेटासेट के भीतर एक चर को संदर्भित करने का प्रयास कर सकते हैं।

आपने एक प्लॉट बनाया है, लेकिन यह आपके प्लॉट विंडो में फिट नहीं है। प्लॉट विंडो का आकार बढ़ाने की कोशिश करें और अपने प्लॉट कमांड को फिर से रन करें।

Error: unexpected numeric constant in: ...

आपको सबसे अधिक संभावना है कि एक कोष्ठक, एक अल्पविराम याद आ रहा है, या जब आप पिछली पंक्ति को पूरा कर चुके थे, तो आपको संकेत के साथ कोड की एक पंक्ति भागा। अपनी कोड लाइन को ध्यान से पढ़ें और सभी उचित सिंटैक्स की जाँच करें।

Error in ....: undefined columns selected

इसका अर्थ है कि आपके द्वारा चयनित डेटा का कॉलम मौजूद नहीं है। यदि संख्यात्मक रूप से कॉलम का चयन करना है, तो सुनिश्चित करें कि आपके पास सूचकांक सही हैं। यदि नाम से चयन किया जाता है, तो वर्तनी और पूंजीकरण की जांच करें। अंत में, यह सुनिश्चित करने के लिए जांचें कि आपने डेटा को सही ढंग से लोड किया है और यह कि चर नाम कॉलम कॉलम के रूप में दिखाई दे रहे हैं और डेटा की पहली पंक्ति के रूप में नहीं।

संदर्भ

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

एम.एस. एक्सेल के द्वारा जैवमितिय विश्लेषण

सुनील कुमार यादव

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, लाईब्रेरी एवेन्यू, नई दिल्ली – 110012

माइक्रोसॉफ्ट एक्सैलसदैव उपयोग किये जाने वाला एक सॉफ्टवेयर अनुप्रयोग है, विश्वभर में लाखों लोगमाइक्रोसॉफ्टएक्सैलका प्रयोग करते हैं। प्रयोक्ता एक्सैलमें हर प्रकार के आंकड़ों की प्रवृष्टि कर सकते हैं तथा वित्तीय, गणतीय एवं सांख्यिकीय गणनायें कर सकते हैं ।

माइक्रोसॉफ्ट एक्सैलआंकड़ों के विश्लेषण के लिए एनालिसिस टूल पैक प्रदान करता है जिसकी सहायता से जटिल सांख्यिकीय अथवा अभियांत्रिकीय विश्लेषण के विभिन्न चरणों कोबचाया (save) जा सकता है तथा प्रत्येक विश्लेषण के लिए आंकड़े एवं प्राचल प्रदान करता है । इस औजार की सहायता से उपयुक्त सांख्यिकीय अथवा अभियांत्रकीय सूक्ष्म गणनायें की जा सकती हैं तथा प्राप्त परिणामों को तालिका में दर्शाया जाता सकता है । कुछ औजार परिणामों की तालिका के अतिरिक्त चार्ट का भीसृजन करते हैं ।

1.स्वतन्त्र प्रतिदर्श t-जांच

दो जनसंख्या औसतों का सांख्यिकीय परीक्षण द्विप्रतिदर्श t-जांच की सहायता से यह जानने के लिये परीक्षण किया जाता है कि क्या दोनों प्रतिदर्श भिन्न हैं, साथ ही जब दोनों सामान्य वितरणों के प्रसरण अज्ञात हों तथा परीक्षण प्रतिदर्श का आकार छोटा हो तो ऐसी परिस्थिति में भी द्विप्रतिदर्श t-जांच का उपयोग किया जाता है ।

निम्नलिखित उदाहरण की सहायता से एक्सेल मे t-जांच की विधि को समझा जा सकता है। t-जांच को शून्य प्राकल्पनाकी जांच के लिए प्रयोग किया जाता है जिसका अभिप्राय है कि दोनों जनसंख्याओं के औसत बराबर है ।

निम्नलिखत उदाहरण मे 6 महिला विद्यार्थियों तथा 5 पुरुष विद्यार्थियों के अध्ययन के घंटों को लिया गया है ।

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

H15 f_x

	A	B	C
1	Female	Male	
2	26	23	
3	25	30	
4	43	18	
5	34	25	
6	18	28	
7	52		
8			
9			

t-जांच के लिए निम्नलिखित चरणों का पालन करें ।

1.सर्वप्रथम यह जानने के लिए कि क्या दोनों जनसंख्याओं के प्रसरण समान है। F-जांच करें ।

(क) डाटा टैब पर 'डाटा एनालिसिस बटन' को क्लिक करें ।

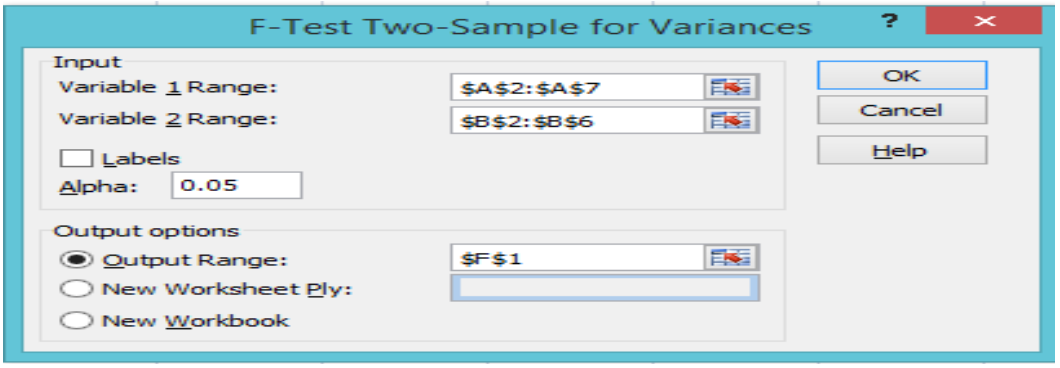
नोट– यदि डाटा एनालिसिस बटन न हो तो 'एनालिसिस टूल पैक' को लोड करें ।

(ख) प्रसरण के लिए द्वि प्रतिदर्श F-जांच को चुनें

(ग) वैरीयेबल–1 रेंज बाक्स को क्लिक करें तथा A_2:A_7 रेंज को चुनें ।

(घ) वैरीयेबल–2 रेंज बाक्स को क्लिक करें तथा B_2:B_6 रेंज को चुनें ।

(च) आउटपुट रेंजबाक्स को क्लिक करें तथा F_1क्लिक करें ।

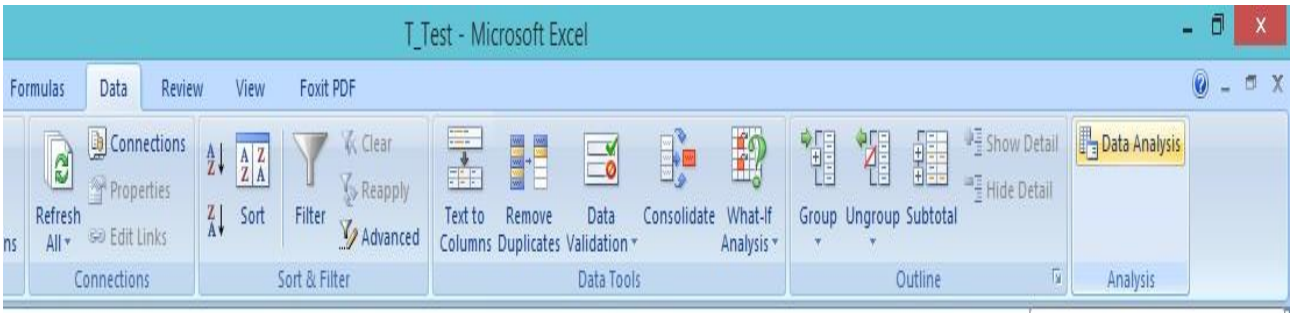


(छ) ओके क्लिक करें ।

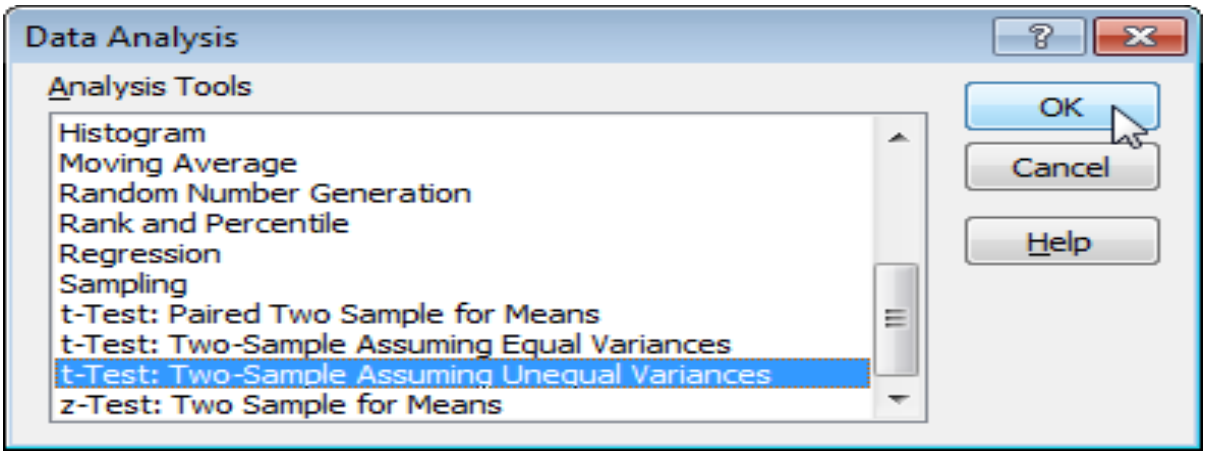
F	G	H
F-Test Two-Sample for Variances		
	Variable 1	*Variable 2*
Mean	33	24.8
Variance	160	21.7
Observations	6	5
df	5	4
F	7.373272	
P(F<=f) one-tail	0.037888	
F Critical one-tail	6.256057	

F (7.373272) Fक्रांतिक(6.256057)से बड़ा है । अतः दोनों जनसंख्याओं के प्रसरण भिन्न हैं ।यह एक भिन्न प्रसरणों की अवस्था है

2.डाटा टैब पर डाटा एनालिसिस क्लिक करें



3.t-टेस्ट : टू सैंपल एज्यूमिंग अनईक्वल वैरीयेसेंस को चुनें तथा ओके क्लिक करें ।

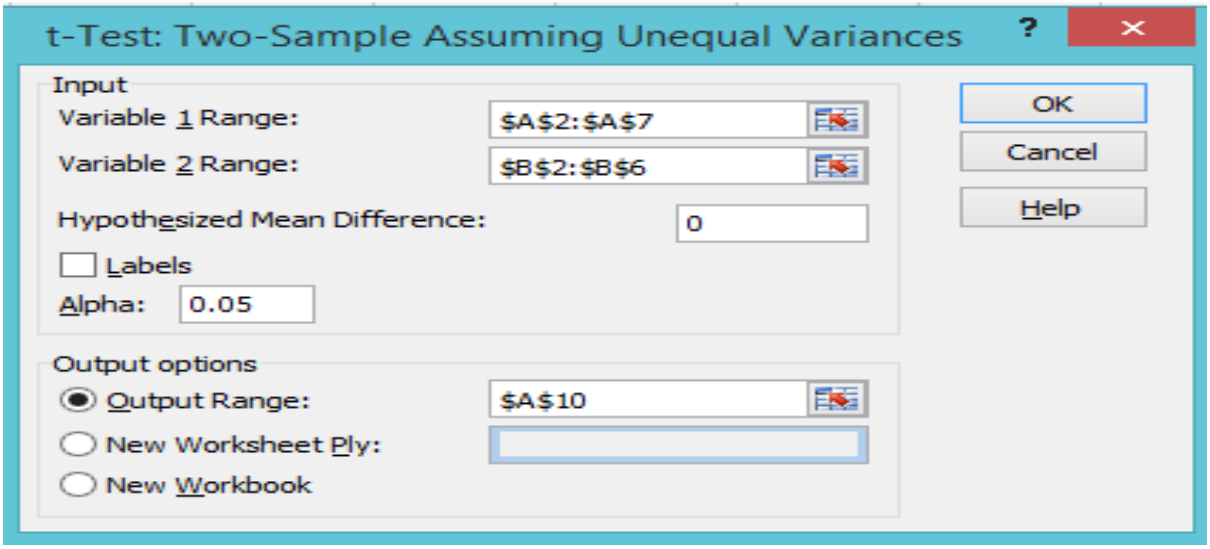


4.वैरीयेबल –1 रेंज बाक्स को क्लिक करें तथा रेंजA2:A7 को चुनें ।

5.वैरीयेबल–2 रेंज को क्लिक करें तथा रेंज B2:B6 को चुनें ।

6.हाइपोथाइज्ड मीन डिफरेंस बाक्स को क्लिक करें तथा '0' टाइप करें ($H_0 : \mu_1 - \mu_2 = 0$)।

7.आउटपुट रेंजबाक्स को क्लिक करें तथा E10 को चुनें ।



8.ओके क्लिक करें

परिणाम:

t-जांच: असमान प्रसरण मानते हुए प्रतिदर्श

	पुरुष	महिला
औसत	33	24-8
प्रसरण	160	21-7
प्रेक्षण	6	5
प्राकल्पित औसत		
अन्तर	0	
स्वतंत्रता की कोटि	7	

t-सांख्यिकी	1.472605	
एकल पुच्छ	0.09217	
क्रान्तिक एकल पुच्छ	1.894579	
क्रान्तिक द्वि. पुच्छ	0.18434	
t- क्रान्तिक द्वि. पुच्छ	2.364624	

निष्कर्षः हम एक द्वि. पुच्छ जांच (असमानता) करते हैं । यदि t- सांख्यिकी <t क्रान्तिक द्वि. पुच्छ अथवा t- सांख्यिकी >tक्रान्तिक द्वि. पुच्छ हों तो हम शून्य प्राकल्पनाको नकार देते हैं परन्तु यहां ऐसा नहीं है । यहां $-2.365 < 1.473 < 2.365$ है ।इसलिए शून्य प्राकल्पनाको नकारा नहीं जा सकता है । प्रतिदर्श औसतों (3.3-24.8) के बीच के अंतर से स्पष्ट नहीं होता है कि महिला और पुरुष विद्यार्थियों के अध्ययन के घंटों में कोई महत्वपूर्ण अंतर है ।

2.युगल t-जांचः.युगल t-जांच प्रायः प्रतिदर्श समूहों के प्राप्ताकों की हस्तक्षेप से पूर्व अथवा पश्चात तुलना के लिए किया जाता है । युगल t-जांच को दो जनसंख्या औसतों की तुलना करने के लिए किया जाता हे । जब हमारे पास दो प्रतिदर्श हों तथा एक प्रतिदर्श के प्रेक्षणों को दूसरे प्रतिदर्श के प्रेक्षणों के साथ जोड़ा बनाया जा सके ।

एक्सैल में युगल t-जांच :

दो युगल मानों (जैसे किसी स्थिति से पूर्व अथवा पश्चात) की तुलना के लिए, जब दोनों प्रेक्षण एक ही वस्तु अथवा अनुरुप वस्तुओं से लिये गये हों, ऐसी अवस्था में युगल t-जांच का प्रयोग किया जा सकताहै । उदाहरण के लिए हमारे पास ८ वस्तुओं के आंकड़ों के दो चर पूर्व और पश्चात हों (आहार से पूर्व एवं पश्चात आर) ।

जांच की प्राकल्पनाइस प्रकार हैः

H_0 : m loss = 0भार में औसत कमी शून्य थी)

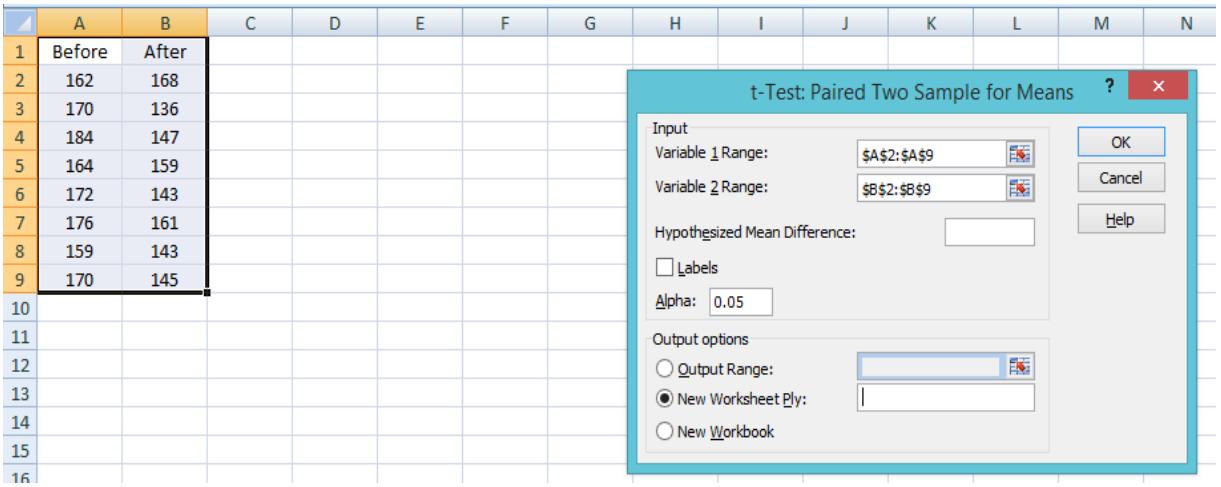
H_a : m loss $\neq$ 0 (भार में औसत कमी शून्य से भिन्न थी)

उदाहरण के लिए भार में कमी के निम्नलिखित आंकड़ों को युगल t-जांच के लिए गया है।

DIET.XLS

पूर्व	पश्चात
162	168
170	136
184	147
164	159
172	143
176	161
159	143
170	145

1. युगल t-जांच के लिए टूल्स डाटा एनालिसिसि /t-टेस्ट: पेर्यड टू सैम्पल फार मीन्सको चुनें ।
2. t-टेस्ट: पेर्यड टू सैम्पल फार मीन्सडायलॉग बाक्स मेंचर –1 की इनपुट रेंज के लिए समूह पूर्वमें भार के 38 मानों को हाइलाइट करें (162 से 170 के मान) । चर–2 की इनपुट रेंज के लिए समूह पश्चातमें भार के 8 मानों को हाइलाइट करें (168 से 145 के मान) । अब अन्य वस्तुओं की उनकी डिफाल्ट अवस्था में छोड़ दें । यहां डायलॉग बाक्स दिखाया गया है । ओके क्लिक करें ।



3. परिणामों को निम्नलिखित आउटपुट तालिका में दशार्या गया है ।

t-जांचः औसतों के लिए युग्मिक द्वि–प्रतिदर्श

	चर–1	चर–2
औसत	169.625	150.25
प्रसरण	65.125	121.9286
प्रेक्षण	8	8
पियरसन सहसंबंध	-0.17675	
परिकाल्पित औसत		
अंतर	0	
स्वतंत्रता की कोटि	7	
t-सांख्यिकी	3.706873	
P(t< = t)एकल पुच्छ	0.003793	
tक्रान्तिक एकल पुच्छ	1.894579	
P(t< = t)द्वि पुच्छ	0.007586	
द्वि पुच्छ	2.364624	

अतः इस t जांच के लिए द्वि पुच्छ P मान है P = 0-008 (0-00758तथा t = 3.71
इस जांच के परिणामों से हम वह प्राप्त नहीं कर सके जो हम वास्तव में चाहते हैं । इसे भली भांति समझने के लिए हमें यह जानना आवश्यक है कि युगल t-जांच वास्तव में दो मानों के बीच के अन्तर की जांच है । अतः एक बेहतर विश्लेषण के लिए पहले पूर्व एवंपश्चात के मानों का अंतर निकालना चाहिए । इसके लिए एक अतिरिक्त स्तम्भ अन्तर का सृजन सूत्र

=A2-B2के प्रयोग से किया गया है तथा सूत्र को शेष सभी सेल में कापी किया गया है । औसत अन्तर की भी गणना की गई है ।

पूर्व	पश्चात	अन्तर
162	168	&6
170	136	34
184	147	37
164	159	5
172	143	29
176	161	15
159	143	16
170	145	25

औसत अन्तर = 19.375
यदि हम मूल प्राकल्पना को देखें तो इसमें औसत का मान शून्य से भिन्न है ।
इस प्रकार t-जांच वास्तव में यह जांच कर रहीं है कि क्या 19.38,इसकी उपयोगिता के दावे के लिए प्रर्याप्त मात्रा में शून्य से भिन्न हे । अतः हम औसत अंतर (कमी) को जानने में अधिक रुचिकर है बजाय पूर्व और पश्चात के व्यक्तिगत औसतों को जानने में ।

इसलिए इन परिणामों को उचित ढंग से दर्शाने के लिए हमें औसत अंतर के मानक विचलन की आवश्यकता है । इसकी गणना अंतर मानों पर वर्णनात्मक सांख्यिकी (टूल्स/डाटा एनालिसिस/डिसक्रिप्टिव स्टेस्टिक्स)की सहायता से की जा सकती है । सारांश सांख्यिकी तथा 95%प्रतिशत कानफिडेंस इन्टरवल ऑप्सन्स को चुनिये ।

परिणाम निम्नलिखित हैं :

स्तम्भ–1	
औसत	19.375
मानक त्रुटि	5.22677
माध्य	20.5
बहुलक	लागू नहीं
मानक विचलन	14.78356
प्रतिदर्श प्रसरण	218.5356
क्रुटोसिस	-0.57529
स्क्यूनैस	43
परास	-6
न्यूनतम	37
अधिकतम	155
योग	8
काउंट	–
कानफिडेंस स्तर (95%)	12.35936

यहां पर यह ध्यान देने योग्य है कि औसत को मानक त्रुटि से भाग करने पर वही मान प्राप्त होता है जो पिछली तालिकी की t-सांख्यिकी से प्राप्त हुई थी । अन्य महत्वपूर्ण सूचना 95 प्रतिशत कानफिडेंस इन्टरवल है । उपरोक्त तालिका से प्राप्त कानफिडेंस स्तर (95%)मान 12.259 है,कानफिडेंस इन्टरवल इस मान से थोड़ी अधिक अथवा थोड़ी कम औसत मान के बराबर है । अतः 95 प्रतिशत कानफिडेंस इन्टरवल पर औसत अंतर (7.01,31.74) है ।

इसे सही ढंग से इस प्रकार व्यक्त किया जा सकता है कि औसत भार हानिशून्य से अधिक है, द्वि पुच्छ p=0.008 इस बात का प्रमाण हैं कि आहार भार को कम करने मेंसक्षमहै। औसत भार हानि के आस पास95 प्रतिशत इन्टरवल(7.01,31.74)है ।

नोटः इस जांच को एकल पुच्छ जांच की तरह भी किया जा सकता है । इसके लिए एक्सेल तालिका से उचित t-सांख्यिकी तथा Pमान को प्रयोग करें ।

नोटः इस जांच को अंतर के शून्य के अतिरिक्त अन्य परिकाल्पित मानों के साथ भी किया जा सकता है । यद्यपि प्रायः शून्य मान को ही प्रयोग किया जाता है ।

एक्सैलयुगल t-जांच डायलाग बाक्स में अन्य परिकाल्पित मानों की प्रवष्ठि का अवसर प्रदान करती है ।

जैवमितिय विश्लेषणः

उदाहरण1: समंजन की शुष्टता के लिए काई–वर्ग के परीक्षण पर विचार करें जब डेटा की दो श्रेणियां हैं (अर्थात् $k=2$)। इस विश्लेषण वर्गों के किसी भी बड़ी संख्या के लिए आसानी से बढ़ाया जा सकता है। यहां $k = 4$

***k =2* के लिए काई–वर्ग समंजन की शुष्टता का परीक्षणः**

H_0:प्रतिदर्श एक आबादी से 9:3:3:1के अनुपात मे पीली–चिकनी : पीली–झुर्रिया : हरी–चिकनी : हरी–झुर्रिया बीज आते हैं।

H_1:प्रतिदर्श एक आबादी से उपर्युक्त चार समलक्षणी के बीज 9:3:3:1 के अनुपात मे नहीं आते हैं।

प्रतिदर्श आंकणों प्रेक्षित मानों f_i के रूप मे, कोष्टक मे प्रत्याषित मानोंमानों F_i के साथ मे अभिलिखित किया गया ।

	पीली–चिकनी	पीली–झुर्रिया	हरी–चिकनी	हरी–झुर्रिया	n
f_i	152	39	53	6	250
(F_i)	140.625	46.875	46.875	15.625	

$v = k$-1=3

χ^2=8.972

0.025< P <0.05

अतः H_0को अस्वीकार कर सकते हैं।

उदाहरण 2द्विप्रतिदर्श t-जांच

H_0: $\mu_d \leq 5$

H_1: $\mu_d \geq 5$

भूखंड (j)	नए उर्वरक के साथ (X_{1j})	पुराने उर्वरक के साथ (X_{2j})	अन्तर (सेमी) d_j
1	67.4	60.6	6.8
2	72.8	66.6	6.2
3	68.4	64.9	3.5
4	66.0	61.8	4.2
5	70.8	61.7	9.1
6	69.6	67.2	2.4
7	67.2	62.4	4.8
8	68.9	61.3	7.6
9	62.6	56.7	5.9

N=9n d=5.611 bu/acre

v=n-1 s_d=0.701bu/ acre

t=d-5/0.701

=0.872

$T_{0.05(1),8}$=1.860

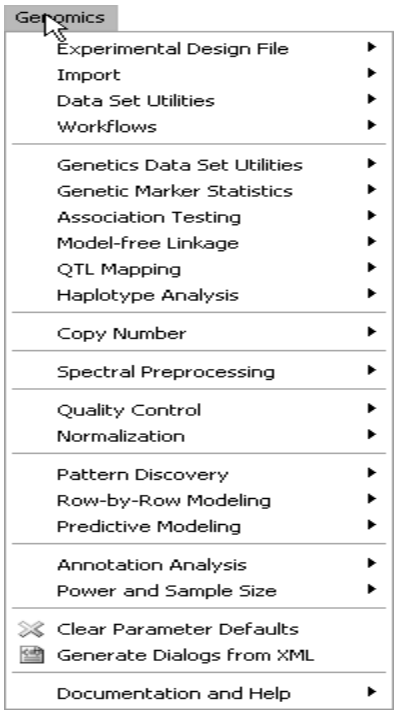
अतः H_0को अस्वीकार नहीं कर सकते है

जे.म.पी-जीनोमिक: ओवरव्यू

डॉ. सुकांत दाश
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

जेएमपी जीनोमिक्स आनुवंशिक मार्कर, माइक्रोएरे और वर्णक्रमीय (प्रोटिओमिक्स और मेटाबॉलिकमिक्स, उदाहरण के लिए) डेटा के एकीकृत सांख्यिकीय विश्लेषण के लिए एक शक्तिशाली डेस्कटॉप सॉफ्टवेयर प्रणाली है। जेएमपी जीनोमिक्स में 100 से अधिक स्वतंत्र विश्लेषणात्मक प्रक्रियाएं (एपी) हैं। हमें आधुनिक जीनोमिक्स विश्लेषण और मानक जेएमपी कार्यक्षमता से जुड़ी शब्दावली और प्रौद्योगिकी से परिचित होना चाहिए। इस सत्र में जेएमपी जीनोमिक्स प्रणाली के प्राथमिक कार्यात्मक पहलुओं का अवलोकन प्रदान करता है, मानक जेएमपी कार्यक्षमता और जेएमपी जीनोमिक्स कार्यक्षमता के बीच कुछ महत्वपूर्ण अंतरों का विवरण, और शामिल नमूना डेटा सेटों का विवरण।

जेएमपी जीनोमिक्स जेएमपी प्लस का पूरी तरह कार्यात्मक संस्करण है और जीनोमिक्स मुख्य मेनू (चित्रा 1.1) में विश्लेषणात्मक प्रक्रिया संवादों का एक संग्रह है। यह 100 से अधिक विश्लेषणात्मक प्रक्रियाओं तक पहुंच प्रदान करता है।



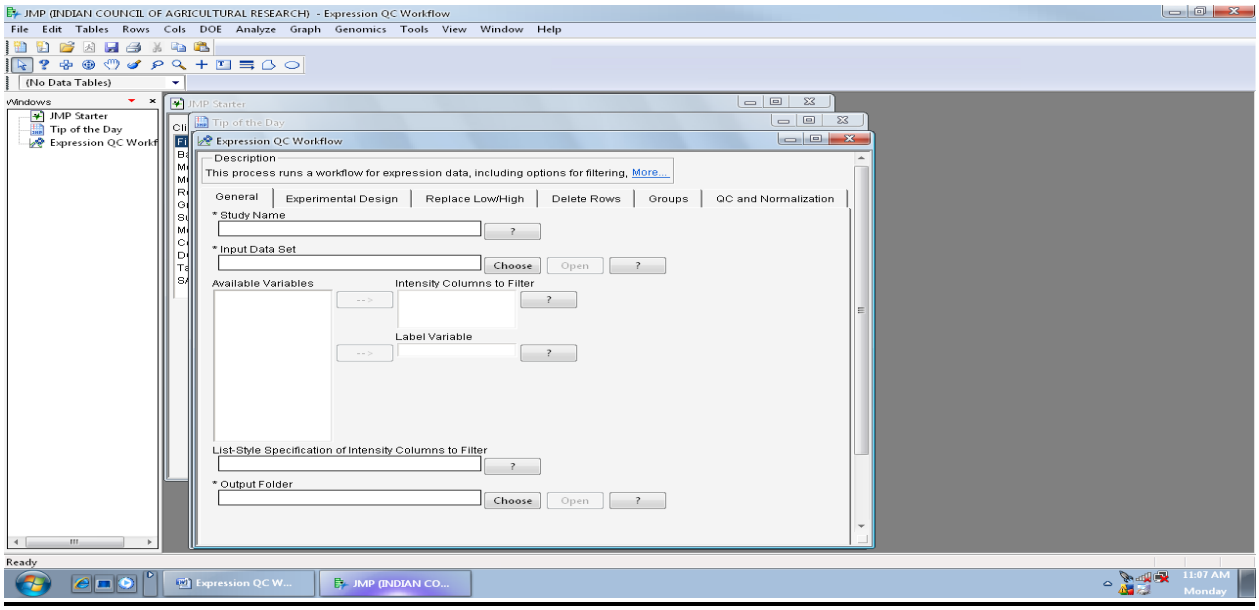
चित्र 1.1 JMP जीनोमिक्स मुख्य मेनू सबमेनस में आयोजित किया गया है।

JMP जीनोमिक्स संवाद मानक JMP संवादों से अलग कार्य करते हैं। मानक JMP संवाद संकलित कोड में गणना को आमंत्रित करते हैं, जबकि JMP जीनोमिक्स संवाद एक SAS प्रोग्राम (प्रत्यय .sas के साथ) उत्पन्न करते हैं, इसे पृष्ठभूमि में निष्पादित करते हैं, और फिर परिणाम वापस करते हैं। JMP जीनोमिक्स संवाद मानक JMP संवादों से अलग कार्य करते हैं। मानक JMP संवाद संकलित कोड में गणना को आमंत्रित करते हैं, जबकि JMP जीनोमिक्स संवाद एक SAS प्रोग्राम (प्रत्यय .sas के साथ) उत्पन्न करते हैं, इसे पृष्ठभूमि में निष्पादित करते हैं, और फिर परिणाम देते हैं। JMP जीनोमिक्स संवाद मानक JMP संवादों से अलग कार्य करते हैं। JStandard JMP संवाद संकलित कोड में गणना को आमंत्रित करते हैं, जबकि JMP जीनोमिक्स संवाद एक SAS प्रोग्राम (प्रत्यय .sas के साथ) उत्पन्न करते हैं, इसे पृष्ठभूमि में निष्पादित करते हैं, और फिर परिणाम वापस करते हैं। परिणाम आम तौर पर एसएएस डेटा सेट (जिसे एसएएस डेटा टेबल के रूप में भी जाना जाता है, प्रत्यय .sas7bdat के साथ जाना जाता है) के साथ-साथ एक JMP स्क्रिप्टिंग भाषा फ़ाइल (प्रत्यय .jsl) के साथ होती है जो स्वचालित रूप से मानक JMP प्लेटफार्मों को आमंत्रित करती है।छोटे

जावा प्रोग्राम कुछ गणनाओं की सुविधा प्रदान करते हैं। अधिकांश JMP जीनोमिक्स संवादों का एक महत्वपूर्ण अंतर यह है कि वे खुले JMP डेटा तालिकाओं को संसाधित नहीं करते हैं। इसके बजाय, वे हमें एक या अधिक एसएएस डेटा सेट निर्दिष्ट करने के लिए संकेत देते हैं जो हमारे फाइल सिस्टम में बनाए और सहेजे गए हैं। यह विशेषता हमें JMP डेटा टेबल के रूप में खोलने और एक प्रक्रिया में कई एसएएस डेटा सेट को निर्दिष्ट किए बिना बहुत बड़े डेटा सेट के साथ काम करने में सक्षम बनाती है। जेएमपी जीनोमिक्स का उपयोग करने में एक प्रारंभिक चुनौती यह तय कर रही है कि किन प्रक्रियाओं को चलाना है और उन्हें किस क्रम में चलाना चाहिए। सॉफ्टवेयर वर्कफ़्लो के निर्माण पर विस्तृत मार्गदर्शन प्रदान नहीं करता है, और आपके खोज उद्देश्यों के आधार पर संभावित वर्कफ़्लो संयोजन की एक विस्तृत विविधता है। मुख्य रूप से JMP जीनोमिक्स द्वारा उपलब्ध कराए गए 8 विभिन्न प्रकार के वर्कफ़्लोज़ हैं। वे हैं बेसिक जेनेटिक्स वर्कफ़्लो, बेसिक कॉपी नंबर वर्कफ़्लो, बेसिक एक्सप्रेशन वर्कफ़्लो, बेसिक miRNA वर्कफ़्लो, बेसिक एक्सॉन वर्कफ़्लो, बेसिक टाइलिंग वर्कफ़्लो, एक्सप्रेशन क्यूसी वर्कफ़्लो और एक्सप्रेशन स्टेटिस्टिक्स वर्कफ़्लो। यहां हमने JMP जीनोमिक्स का उपयोग करते हुए अभिव्यक्ति QC वर्कफ़्लो चलाने की प्रक्रिया के बारे में चर्चा की।

अभिव्यक्ति QC वर्कफ़्लो प्रक्रिया आगे के विश्लेषण के लिए कच्चे अभिव्यक्ति डेटा को साफ करने और तैयार करने के लिए एक वर्कफ़्लो चलाता है।

जेनोमिक्स> वर्कफ़्लोज़> एक्सप्रेशन QC वर्कफ़्लो को सचित्र डायलॉग खोलने के लिए चुनें। संवाद छह टैब से बना है: सामान्य, प्रायोगिक डिजाइन, लो / हाई, रिप्स, पंक्तियों, समूहों और QC और सामान्यीकरण को बदलें 1.1 में दिखाया गया है।



चरण 1

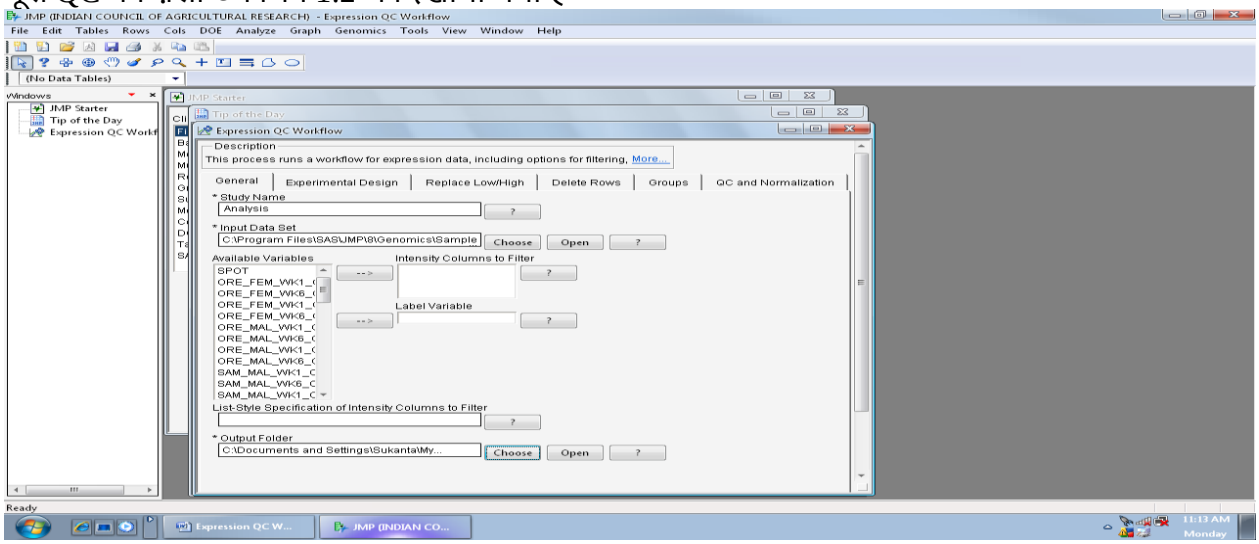
- अध्ययन के नाम के रूप में विश्लेषण (कोई भी नाम जिसे आप असाइन कर सकते हैं) लिखें।..
- Set इनपुट डेटा सेट का चयन करने के लिए चुनें पर क्लिक करें।
- Open एक ओपन डाटा विंडो खुलती है।
- \ नमूना डेटा \ MicroArray \ Scanalyze ड्रोसोफिला फ़ोल्डर में नेविगेट करें।
- drosophilaaging.sas7bdat फ़ाइल का चयन करें और खोलें पर क्लिक करें। 1.2 में दिखाया गया है

नोट: drosophilaaging.sas7bdat फ़ाइल का चयन किया गया है। यदि आप Open पर क्लिक करते हैं, तो आप डेटा सेट देखेंगे।Field उपलब्ध चर चर में सूचीबद्ध चर की जांच करें।

- Set इनपुट डेटा सेट से सभी स्तंभ नाम उपलब्ध चर क्षेत्र में सूचीबद्ध हैं।

- इस उदाहरण में, सभी संख्यात्मक डेटा फ़ील्ड का मूल्यांकन किया जाता है। फ़िल्टर फ़ील्ड और लेबल चर फ़ील्ड को रिक्त करने के लिए तीव्रता कॉलम को छोड़
  - ✓ क्लिक चुनें।
  - ✓ उस फ़ोल्डर पर नेविगेट करें जहां आप आउटपुट रखना चाहते हैं या एक नया फ़ोल्डर बनाना चाहते हैं।
  - ✓ ओके पर क्लिक करें।

पूरा QC वर्कफ़्लो टैब चित्र 1.2 में दिखाया गया है



चरण 2

प्रायोगिक डिजाइन पर क्लिक करें.

- EDDS चुनने के लिए Choose पर क्लिक करें।
- \ नमूना डेटा \ MicroArray \ Scanalyze ड्रोसोफिला निर्देशिका में नेविगेट करें।
- drosophilaaging_exp.sas7bdat फ़ाइल का चयन करें और खोलें पर क्लिक करें.

Drosophilaaging_exp.sas7bdat फ़ाइल का चयन किया गया है। यदि आप Open पर क्लिक करते हैं, तो आप डेटा सेट देखेंगे।

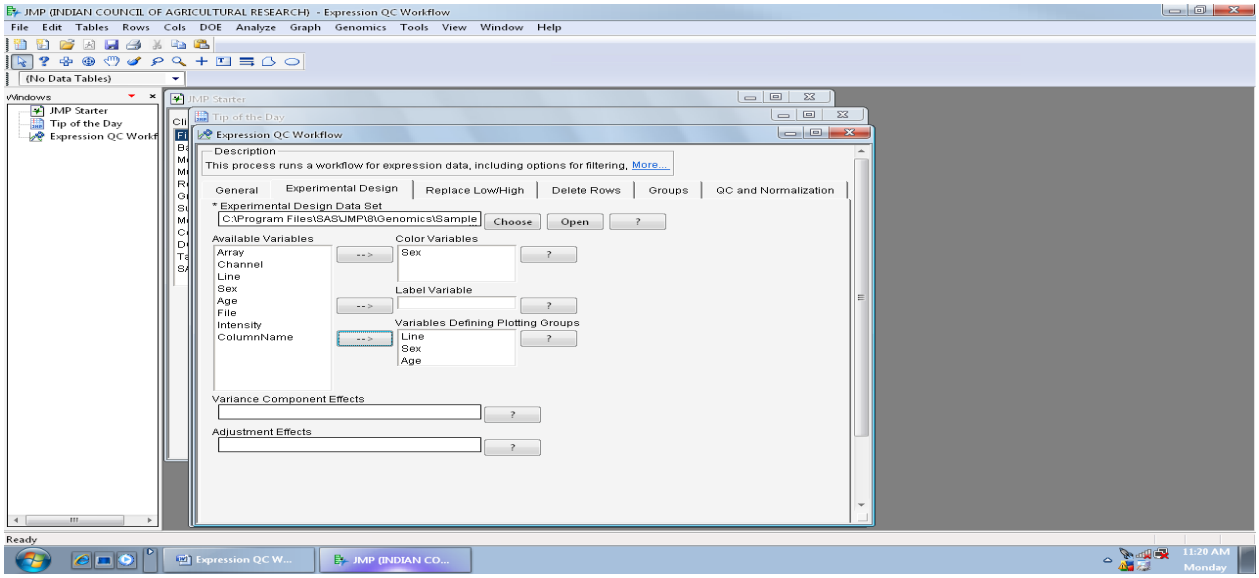
सेक्स द्वारा आउटपुट प्लॉट पर अंकों को अलग-अलग रंग देने के लिए, निम्नलिखित चरणों को पूरा करें:

उपलब्ध चर सूची से सेक्स हाइलाइट करें.

- Ables कलर वेरिएबल्स फ़ील्ड में सेक्स जोड़ने के लिए क्लिक करें।
- लेबल चर क्षेत्र को खाली छोड़ दें।
- प्रत्येक प्रयोगात्मक परिस्थितियों (लिंग, आयु और रेखा) के लिए आउटपुट प्लॉट तैयार करने के लिए, निम्नलिखित चरणों को पूरा करें:
  - ✓ उपलब्ध चर सूची से हाईलाइट लाइन।
  - ✓ प्लॉटिंग समूह फ़ील्ड को परिभाषित करने वाले चर में रेखा जोड़ने के लिए क्लिक करें।
  - ✓ उपलब्ध चर सूची से सेक्स हाइलाइट करें।
  - ✓ प्लॉटिंग ग्रुप फ़ील्ड्स को परिभाषित करने वाले चर में सेक्स जोड़ने के लिए क्लिक करें।

- उपलब्ध चर सूची से light हाइलाइट आयु।
- प्लॉटिंग ग्रुप्स फ़ील्ड को परिभाषित करने वाले चर में आयु जोड़ने के लिए क्लिक करें।

संशोधन के लिए भिन्न घटक प्रभाव और समायोजन प्रभाव क्षेत्र अनुपलब्ध हैं। पूरा प्रयोगात्मक डिज़ाइन टैब चित्र 1.3 में दिखाया गया है



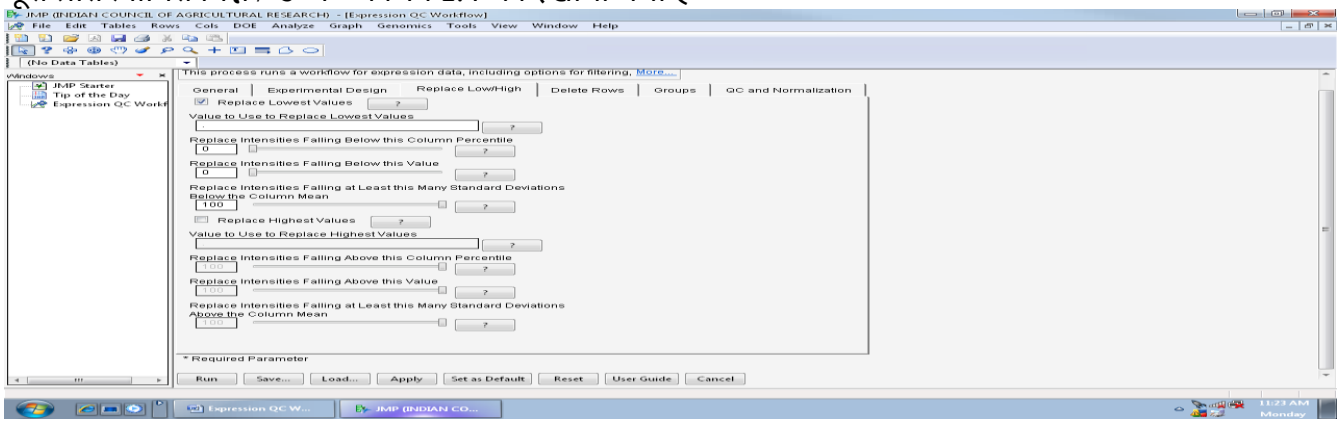
चरण 3

लो / हाई बदलें को क्लिक करें।

इस उदाहरण में, हम निम्न मानों को लापता मान प्रतीक (.) के साथ प्रतिस्थापित करते हैं।

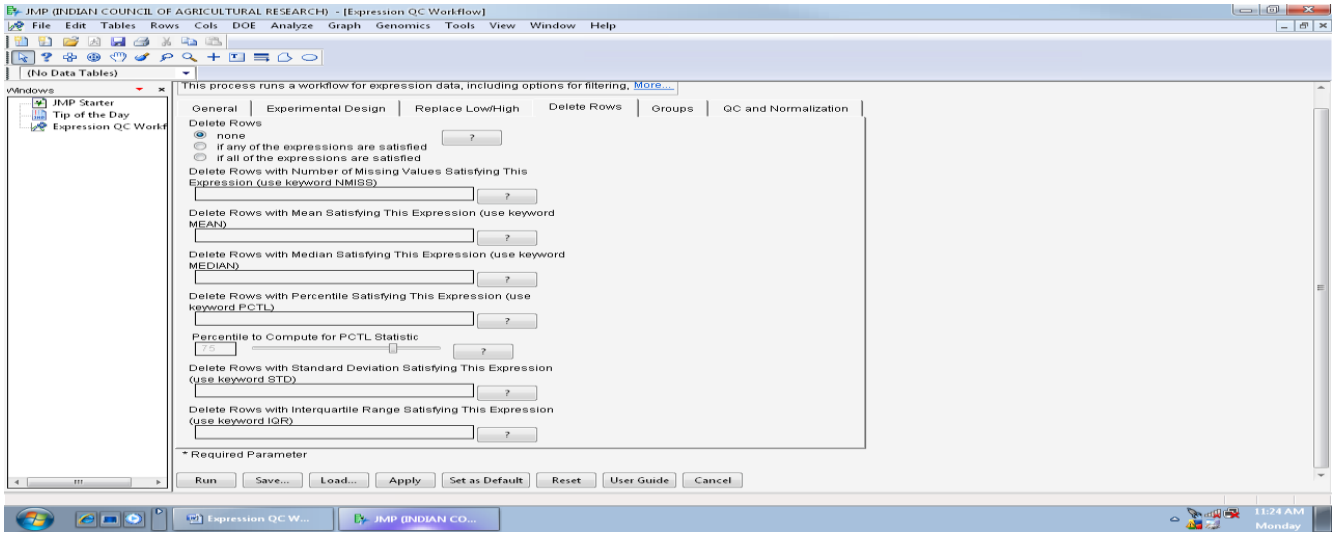
- बदलें निम्नतम मान चेक बॉक्स को चेक करें।
- Field निम्नतम मान फ़ील्ड को प्रतिस्थापित करने के लिए उपयोग करने के लिए मान में एक अवधि (.) लिखें।
- Per इस कॉलम परसेंटाइल क्षेत्र के नीचे आने वाली तीव्रता के प्रकार में 0 टाइप करें।
- इस मान क्षेत्र के नीचे आने वाली तीव्रता तीव्रता में 0 टाइप करें।
- At कम से कम यह कई मानक विचलन नीचे कम से कम प्रतिस्थापन तीव्रता में टाइप करें
- कॉलम मीन फील्ड।
- बदलें उच्चतम मान चेक बॉक्स को अनियंत्रित छोड़ दें.

पूरा प्रतिस्थापित निम्न / उच्च टैब चित्र 1.4 में दिखाया गया है



चरण 4

- हटाएँ पंक्तियों पर क्लिक करें.
- Tab हटाएं पंक्तियों की जाँच करें टैब।
- Tab हटाएं पंक्तियों के टैब में कोई बदलाव न करें।
- पूर्ण हटाए गए टैब को आकृति 1.5 में दिखाया गया है



चरण 5

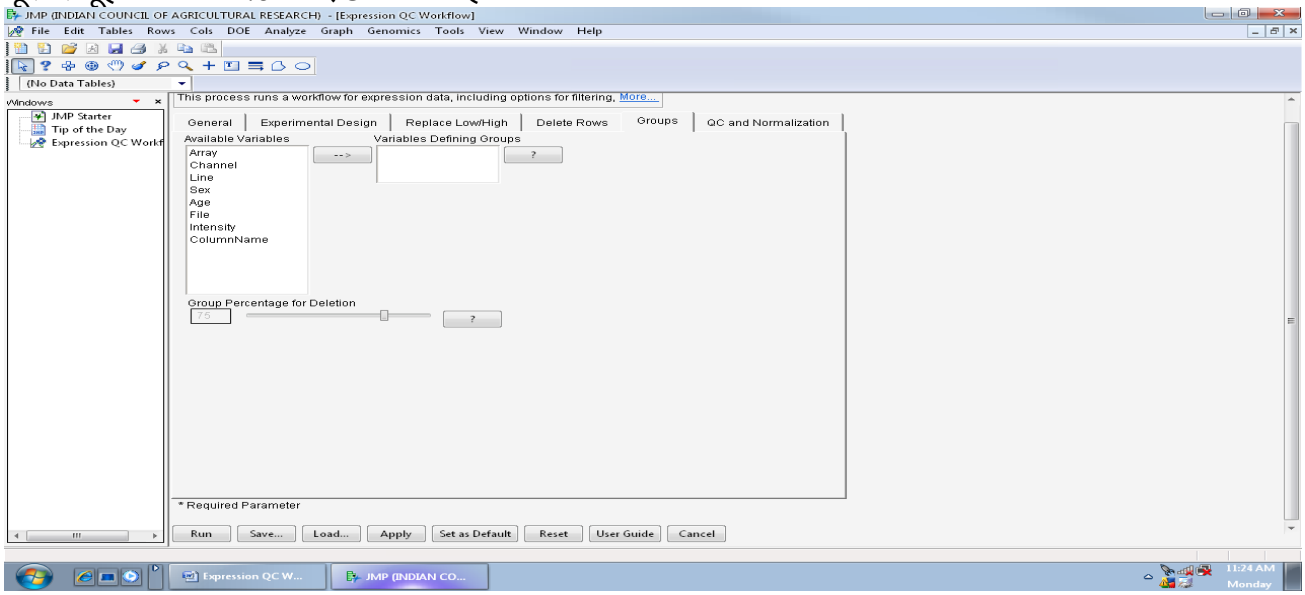
समूह पर क्लिक करें।

- समूह टैब की जांच करें।
- EDDS में सभी कॉलम उपलब्ध वेरिएबल्स फ़ील्ड में सूचीबद्ध हैं।

इस उदाहरण में, किसी समूह को परिभाषित नहीं किया गया है।

- To समूह टैब में कोई परिवर्तन न करें.

पूरा समूह टैब चित्र 1.6 में दिखाया गया है



चरण 6

QC और Normalization पर क्लिक करें.

- क्यूसी और सामान्यकरण टैब की जांच करें।

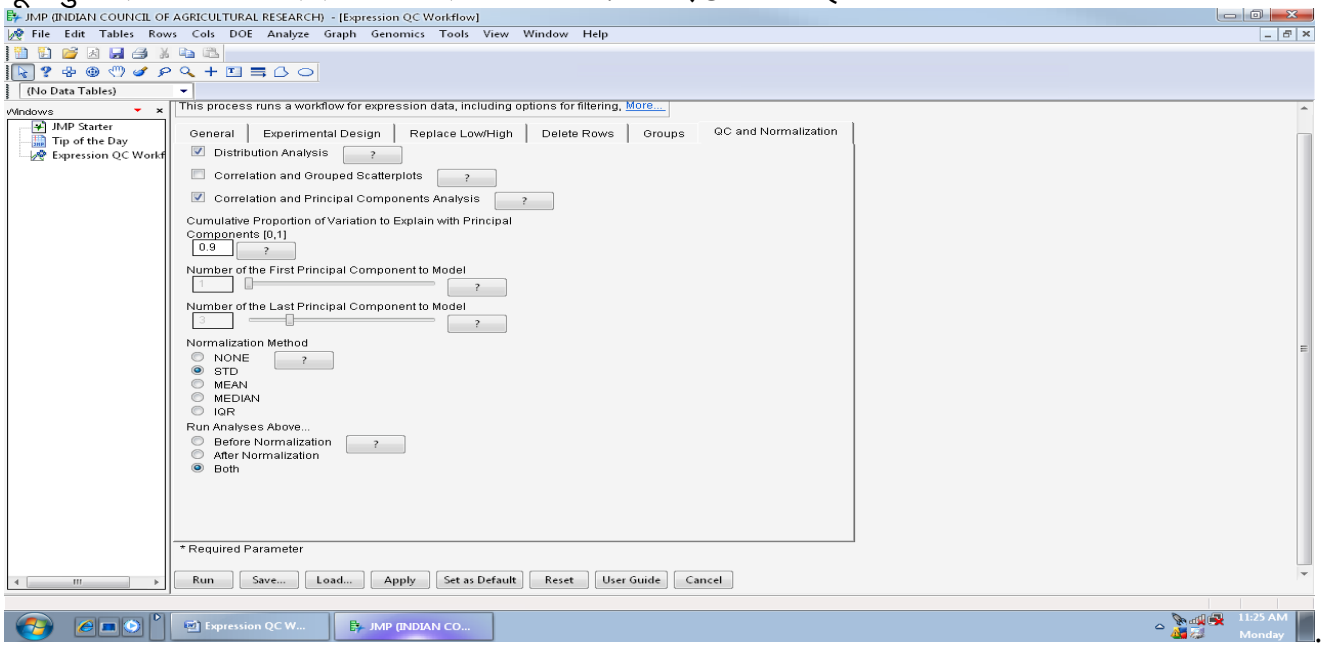
यह उदाहरण सहसंबंध और प्रधान घटक विश्लेषण को छोड़कर सभी डिफ़ॉल्ट गुणवत्ता नियंत्रण और सामान्यीकरण विकल्पों का उपयोग करता है।

सुनिश्चित करें कि वितरण विश्लेषण चेक बॉक्स चेक किया गया है.

- Ation सहसंबंध और प्रमुख घटक विश्लेषण चेक बॉक्स को अनचेक करें

- , मॉडल के पहले और अंतिम प्रमुख घटकों को निर्दिष्ट करने की क्षमता उपलब्ध नहीं है।
- ST सुनिश्चित करें कि एसटीडी को सामान्यीकरण विधि के रूप में चुना गया है।
- The सुनिश्चित करें कि क्यूसी विश्लेषण दोनों सामान्यीकरण किए जाते हैं।

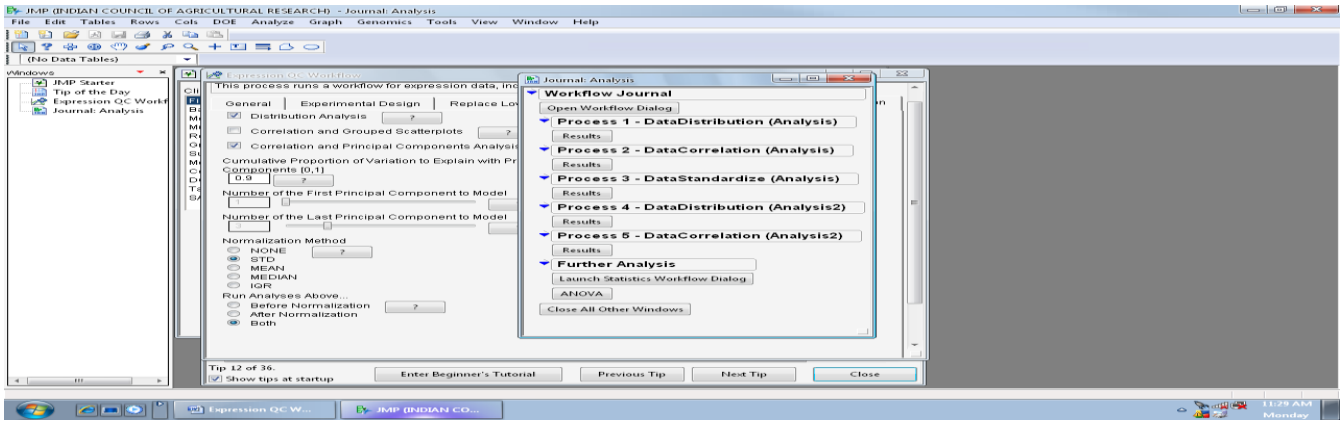
पूर्ण गुणवत्ता नियंत्रण और सामान्यीकरण टैब चित्र 1.7 में दिखाया गया है।



रन पर क्लिक करें

परिणाम

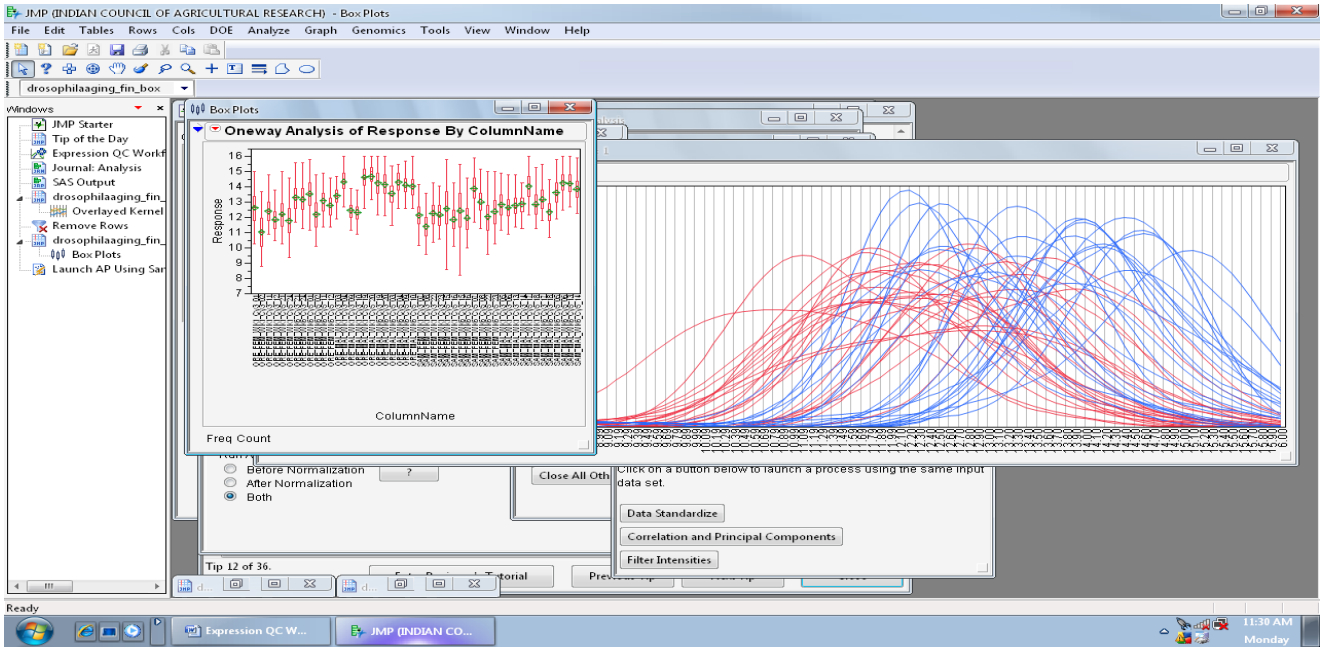
जब आप चलाएँ क्लिक करते हैं, तो एक्सप्रेशन QC वर्कफ़्लो प्रक्रिया वर्कफ़्लो बिल्डर को खोलकर शुरू होती है। वर्कफ़्लो बिल्डर एक्सप्रेशन QC वर्कफ़्लो संवाद में निर्दिष्ट डेटा सेट और मापदंडों से प्रत्येक एपी के लिए सेटिंग्स फ़ाइलों का निर्माण करता है। एक बार सेटिंग फ़ाइल जनरेट और सेव होने के बाद, वर्कफ़्लो में अलग-अलग APs क्रमिक रूप से खोले जाते हैं, आबाद होते हैं और चलते हैं। क्यूसी और सामान्यीकरण एपी के परिणाम निर्दिष्ट आउटपुट फ़ोल्डर में सहेजे जाते हैं, लेकिन प्रदर्शित नहीं होते हैं। इसके बजाय, एक JMP जर्नल उत्पन्न होता है, जो वर्कफ़्लो संवाद और प्रत्येक AP (चित्र 2.1) के परिणामों को लिंक प्रदान करता है।



चित्र 2.1 में दी गई पत्रिका की जाँच करें। ओपन वर्कफ़्लो डायलॉग बटन और नोट करें परिणाम बटन (प्रत्येक प्रक्रिया के तहत स्थित)।

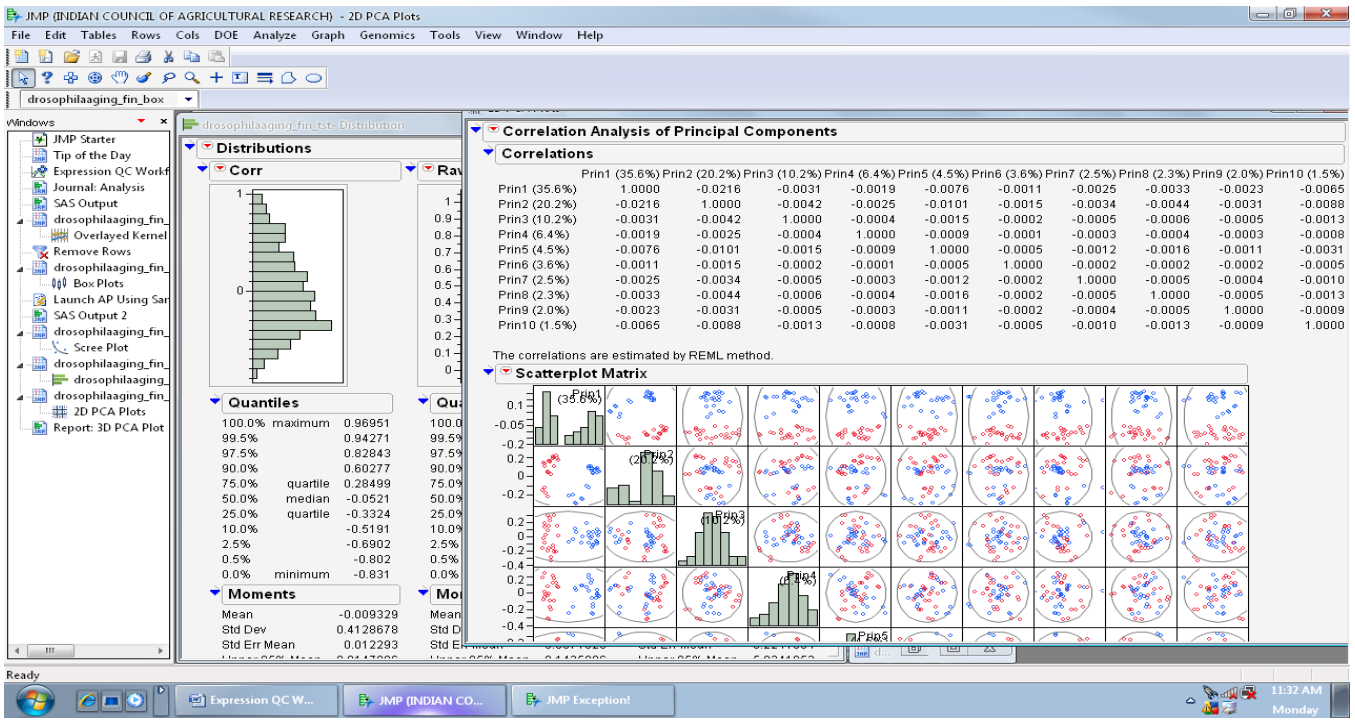
वर्कफ़्लो बिल्डर संवाद, चित्र 2.1 में दिखाया गया है। वर्कफ़्लो बिल्डर संवाद वर्कफ़्लो में प्रत्येक AP के लिए सेटिंग्स दिखाता है। आप अपने विश्लेषण को समायोजित करने के लिए व्यक्तिगत सेटिंग्स का चयन और संपादन कर सकते हैं।

प्रक्रिया 1 के तहत परिणाम पर क्लिक करें - जर्नल विंडो में DataDistribution। डेटा डिस्ट्रिब्यूशन एपी द्वारा उत्पन्न आउटपुट कई अलग-अलग विंडो (चित्र 2.2) में खुलता है।



डेटा वितरण एपी डेटा के दो अलग-अलग भूखंडों को उत्पन्न करता है, एक समानांतर भूखंड जो प्रत्येक कॉलम में टिप्पणियों का एकतरफा वितरण दिखा रहा है और प्रत्येक वितरण के आंकड़े दिखाते हुए एक बॉक्स प्लॉट। इन अतिव्यापी खिड़कियों को चित्र 2.2 में दिखाया गया है

अब प्रक्रिया 2 के परिणाम (डेटा सहसंबंध alnalysis) पर क्लिक करें। डेटा सहसंबंध विश्लेषण द्वारा उत्पन्न आउटपुट आंकड़ा 2.3 में दिखाया गया है



जेएमपी जीनोमिक्स मुख्य रूप से बड़े डेटा सेट के लिए विश्लेषण प्रदान कर रहा है जो अन्य सॉफ्टवेयर का उपयोग करके प्रदर्शन नहीं कर सकता है। यह सॉफ्टवेयर विशेष रूप से आनुवंशिक मार्कर, माइक्रोएरे और वर्णक्रमीय (प्रोटिओमिक्स और मेटाबॉलिकम) डेटा के एकीकृत सांख्यिकीय विश्लेषण के लिए डिज़ाइन किया गया है।

अशोका - एक परिचय

डॉ के. के. चतुर्वेदी
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

जैव सूचना विज्ञान जीव विज्ञान, संगणक विज्ञान, गणित विज्ञान एवं सांख्यिकी विषयों के आपसी सहयोग से मिलकर बना है। जैव सूचना विज्ञान जीन एवं उनके कारकों के कार्य को समझने, कृषि उत्पादकता बढ़ाने, उन्नत किस्मों एवं नस्लों के विकास में सहायक है। जेनेटिक इंजीनियरिंग और जीनोमिक दृष्टिकोण से कृषि सम्बन्धी उत्पादों की उत्पादकता और गुणवत्ता की विशेषताओं को बढ़ाने के लिए, जैव सूचना विज्ञान का एक नए विषय के रूप में सृजन हुआ हैं।

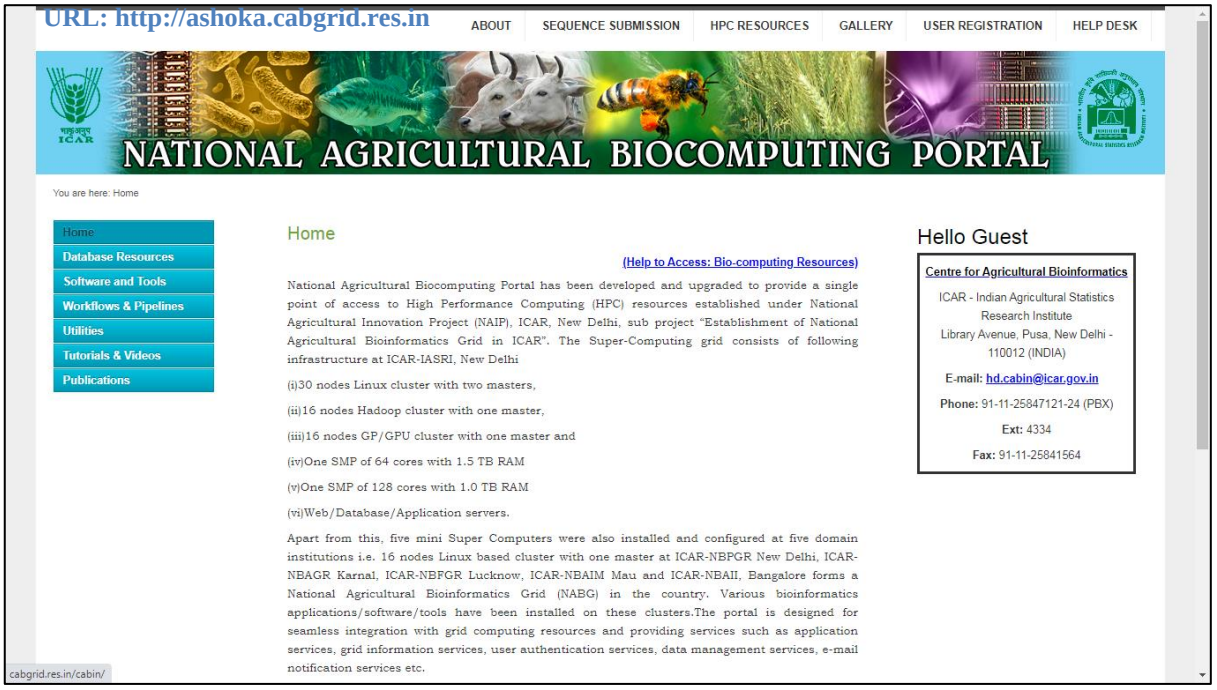
दुनिया भर में आणविक प्रयोगशालाओं के विकास एवं जैविक अनुक्रमण प्रौद्योगिकियों में प्रगति के कारण, बहुत अधिक मात्रा में जैविक आंकड़े उत्पन्न हो रहे हैं। सूचना और संचार प्रौद्योगिकी के क्षेत्र में विकसित नई तकनीकियां इन आंकड़ों को एकत्रित, संग्रहित, संचित एवं विश्लेषित करने में सहायक सिद्ध हो सकती हैं। मूर के नियमानुसार, कंप्यूटर की गणना करने की क्षमता डेढ़ से दो महीनों में दुगुनी हो जाती है। जैविक आंकड़ों में छिपे हुए जैविक ज्ञान को निकालने के लिए उच्च प्रदर्शन कंप्यूटिंग सुविधाओं की जरूरत है। उच्च प्रदर्शन कंप्यूटिंग या हाई परफॉरमेंस कंप्यूटिंग (एचपीसी) आंकड़ों को जल्दी, कुशलतापूर्वक एवं उन्नत एप्लीकेशन सॉफ्टवेयर की सहायता से विश्लेषित कर सकता है। एचपीसी की क्षमता का आंकलन फ्लॉप्स (FLOPS - Floating point operations per second) में किया जाता है। एचपीसी तकनीकी रूप से एक सुपर कंप्यूटर के रूप में सबसे ज्यादा प्रचलित हुआ है।

भारतीय कृषि अनुसंधान परिषद (आई.सी.ए.आर.) ने भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली में कृषि जैव सूचना विज्ञान केंद्र की स्थापना की है। भारतीय कृषि अनुसंधान परिषद (आई.सी.ए.आर.) ने विश्व बैंक द्वारा पोषित राष्ट्रीय कृषि नवोन्मेषी परियोजना (एन.ए.आई.पी.) के अंतर्गत भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली में एक उप परियोजना राष्ट्रीय कृषि जैव सूचना ग्रिड (एन.ए.बी.जी.) की आईसीएआर में स्थापना की स्वीकृत की। इस परियोजना में भा.कृ.अनु.प.–राष्ट्रीय पादप आनुवंशिक संसाधन ब्यूरो (एन.बी.पी.जी.आर.) नई दिल्ली, भा.कृ.अनु.प.–राष्ट्रीय पशु आनुवंशिक संसाधन ब्यूरो (एन.बी.एजी.आर.) करनाल, भा.कृ.अनु.प.–राष्ट्रीय मत्स्य आनुवंशिक संसाधन ब्यूरो (एन.बी.एफ.जी.आर.) लखनऊ, भा.कृ.अनु.प.–राष्ट्रीय कृषि उपयोगी सूक्ष्मजीवों ब्यूरो (एन.बी.ऐ.आइ.एम.) मऊ और भा.कृ.अनु.प.–राष्ट्रीय कृषि कीट संसाधन ब्यूरो (एन.बी.ऐ.आई.आर.), बंगलुरु सहयोगी संस्थान थे। इस परियोजना का प्रमुख उद्देश्य जैव आंकड़ों के विश्लेषण हेतु उच्च प्रदर्शन कंप्यूटिंग या हाई परफॉरमेंस कंप्यूटिंग (एचपीसी) की स्थापना एवं जीनोमिक डेटा संसाधनों और विभिन्न जैविक डेटाबेस के विकास, विश्लेषण और भंडारण करना है ।

संस्थान में एचपीसी की स्थापना मिश्रित रूप में की गयी है । इसमें चार अलग अलग अर्थात् 40 नोड्स लाइनक्स, 16 नोड्स जी पी–जी पी यू लाइनक्स, 16 नोड्स बिग डाटा, एक 1.5 टीबी रैम और एक 1.0 टीबी रैम से निहित सममित बहु–प्रोसेसर (एसएमपी) के रूप में सुपरकंप्यूटर स्थापित किये गए हैं । आंकड़ों को रखने एवं विश्लेषित करने हेतु भंडारण क्षमता को तीन घटकों (अ) नेटवर्क फाइल सिस्टम (ब) समानांतर फाइल सिस्टम और (स) संग्रह प्रणाली (आर्काइवल) में विभाजित किया गया है। इन सभी को जोड़ने के लिए तीन प्रकार के नेटवर्क बनाये गए हैं। क्यू–लॉजिक का उच्च बैंडविड्थ नेटवर्क (क्यूडीआर इनफिनीबैंड स्विच) सभी नोड्स एवं भण्डारण क्षमता प्रणाली के बीच में सम्बन्ध स्थापित कर एक दूसरे को सन्देश पहुँचाने में सहायता करता है। गीगाबिट नेटवर्क का उपयोग क्लस्टर

प्रशासन और प्रबंधन के लिए किया गया है। आईएलओ–3 नेटवर्क का उपयोग समस्त नोड्स अन्य उपकरणों के स्वास्थ्य की निगरानी एवं प्रबंधन के लिए किया गया है। प्रत्येक सहयोगी संस्थान में भी एक 16 नोड्स लाइनक्स सुपरकंप्यूटर स्थापित किया है । इन पांचों संस्थानों के सुपरकंप्यूटर को भी मुख्य संस्थान के साथ एम पी एल एस कनेक्टिविटी के द्वारा एकीकृत किया गया है (चित्र १)। ये सुपर कंप्यूटर उपयोगकर्ताओं के लिए एक मिश्रित वास्तुकला का अनूठा उदाहरण प्रस्तुत करते हैं ।

इस सुपरकंप्यूटर का नाम अशोका (ASHOKA: Advanced Supercomputing Hub for Omics Knowledge in Agriculture) दिया गया है। यह सुविधा जैव सूचना उपकरण, डेटाबेस निर्माण और उनके उपयोग से जैविक अनुसंधान को उन्नत बनाने में सहायक सिद्ध हो रहा है । अशोका को कमांड लाइन इंटरफेस (सी.एल.आई.) और वेब पोर्टल की सहायता से प्रयोग कर सकते हैं। अशोका प्रयोग करने लिए सर्वप्रथम पंजीकरण करना अनिवार्य है। पंजीकरण बायो–कंप्यूटिंग पोर्टल के माध्यम से किया जा सकता है। बायो–कंप्यूटिंग पोर्टल चित्र 1 में दर्शाया गया है।

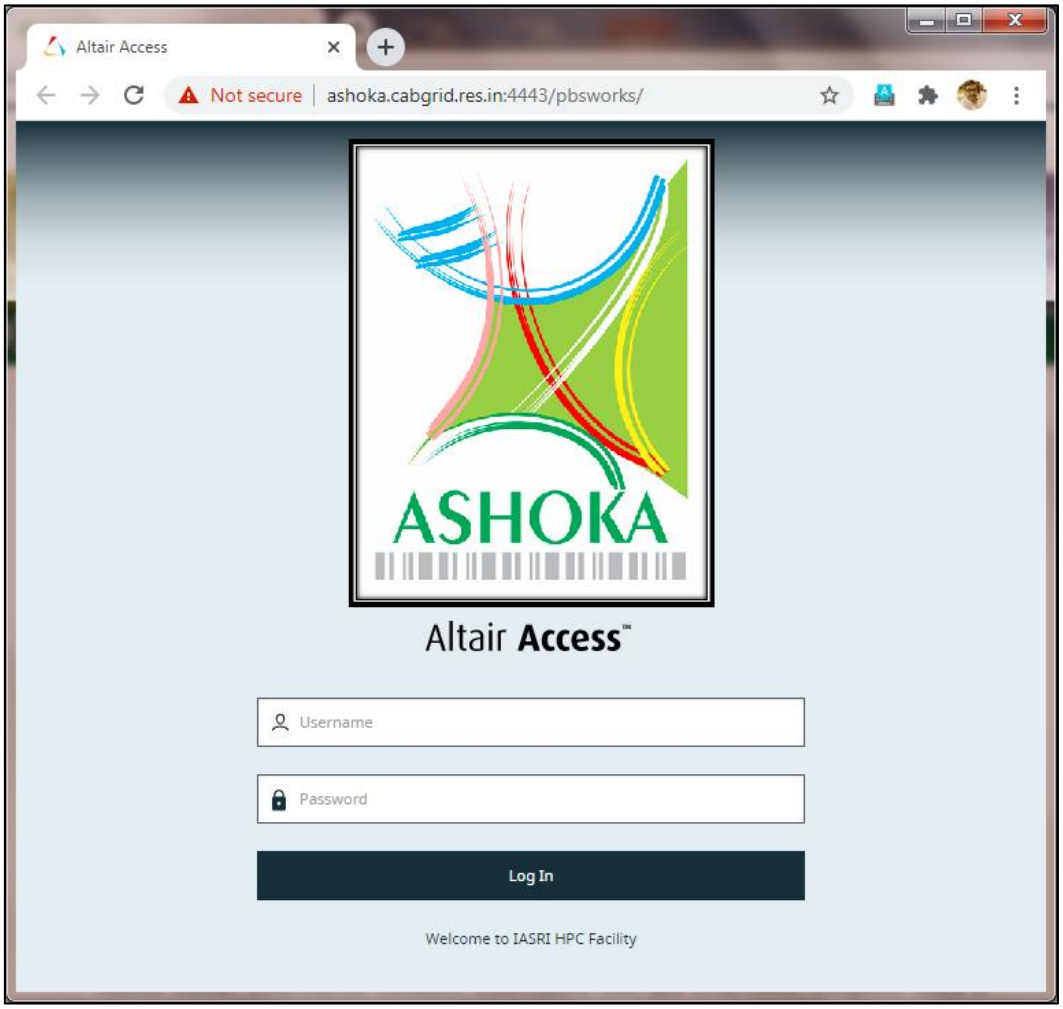


चित्र 1 रू बायो–कंप्यूटिंग पोर्टल

प्रयोगकर्ता को उनकी आवश्यकतानुसार वर्गीकृत किया है जो कि निम्न प्रकार से हैं

1. केंद्र उपयोगकर्ता – प्रभाग में कार्यरत वैज्ञानिकों एवं शोध सहायकों के लिए
2. परिषद् उपयोगकर्ता – परिषद् के संस्थानों में कार्यरत वैज्ञानिकों एवं शोध सहायकों के लिए
3. अन्य उपयोगकर्ता (आर यू) – अन्य संस्थान के वैज्ञानिकों एवं शोध सहायकों के लिए

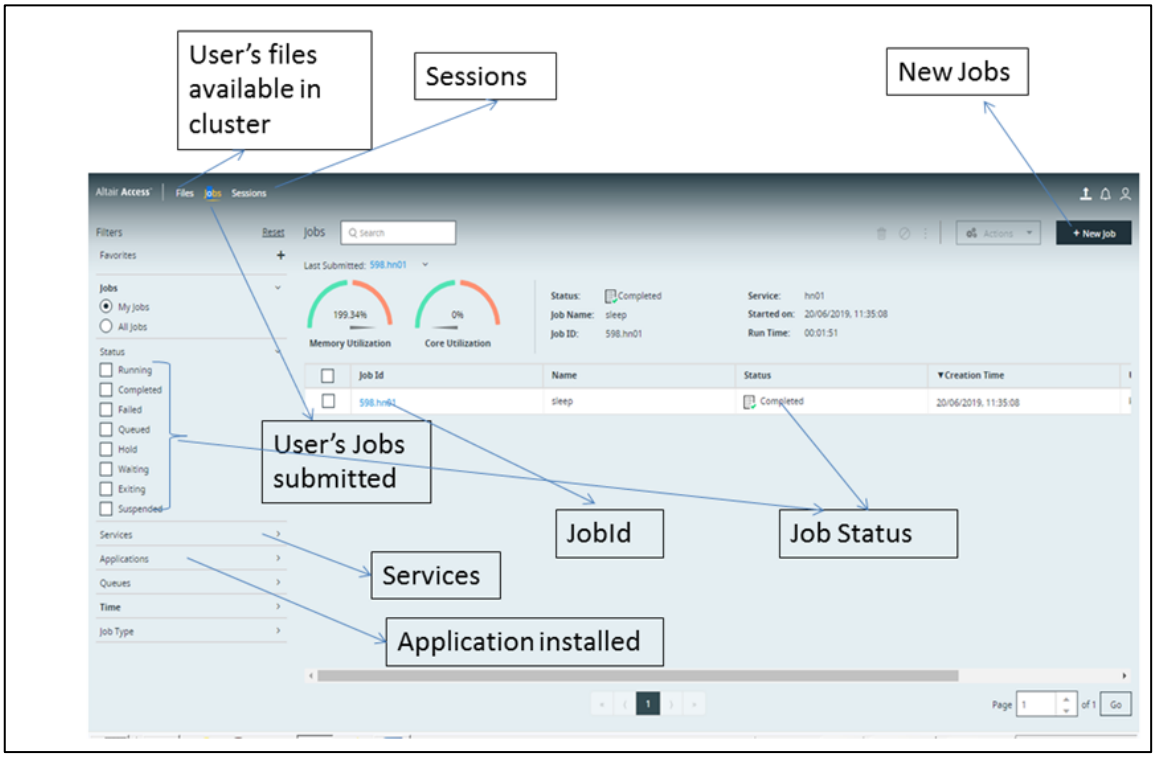
उपयोगकर्ताओं की अनुसंधान गतिविधियों और आवश्यकताओं के अनुसार उनके वर्ग को बदला भी जाता है। इस पोर्टल में विभिन्न मुक्त स्रोत सॉफ्टवेयर का मैत्रीपूर्ण ग्राफिकल यूजर इंटरफेस (जी यू आई) बनाया गया है जोकि जैविक वैज्ञानिकों एवं शोध सहायकों के लिए अत्यंत उपयोगी है । अशोका में लोगिन और प्रवेश करने के लिए वेबपेज को चित्र २ में दर्शाया गया है। पोर्टल प्रमाणीकृत उपयोगकर्ताओं को अपने–अपने स्थानों से उनके जैविक डेटा विश्लेषण एवं प्रदर्शन करने में सक्षम है। इसके विकास के दौरान उपयोगकर्ता की आवश्यकताओं को ध्यान में रख कर निर्मित किया गया है।



चित्र 2 रू अशोका में लोगिन और प्रवेश

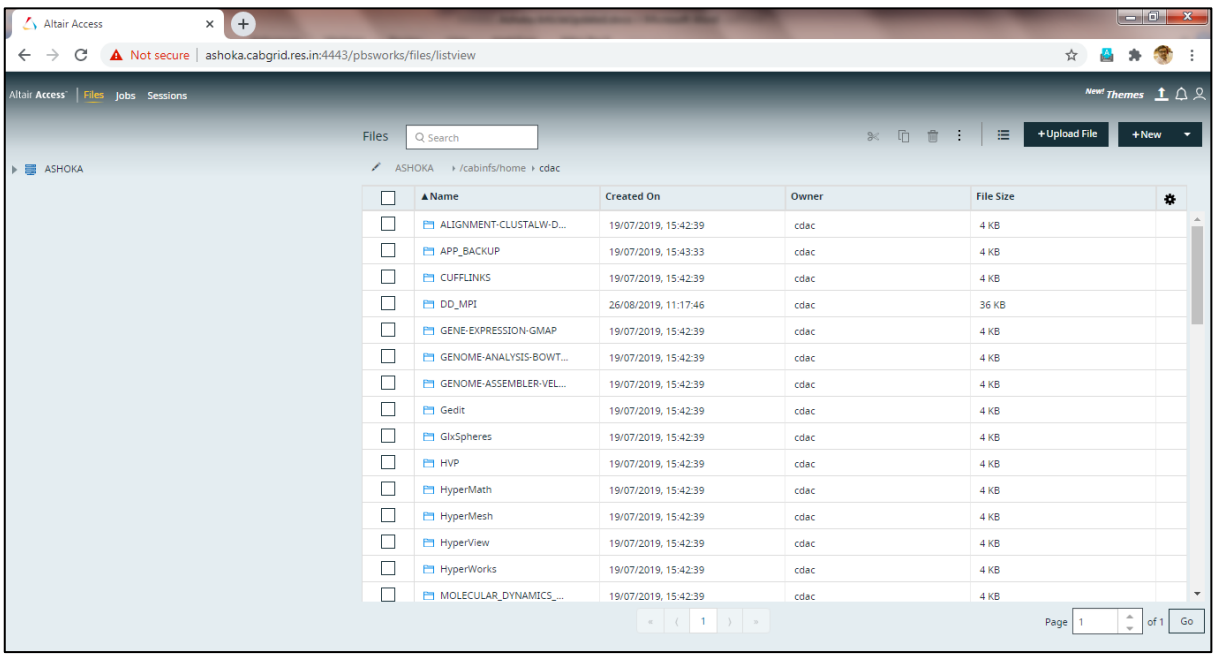
उपयोगकर्ता सफलतापूर्वक लोगिन करने के पश्चात, वह अपनी जॉब प्रस्तुत, निगरानी और प्रबंध कर देख सकता है और जॉब की प्रगति को देखते हुए उचित निर्णय भी ले सकता है कि जॉब को आगे चलना चाहिए या बंद कर देना चाहिए। कई महत्त्वपूर्ण सॉफ्टवेयर को उपयोगकर्ताओं की आवश्यकताओं को ध्यान में रखते हुए अक्सर इस्तेमाल में आने वाले विकल्पों के साथ पोर्टल में एकीकृत किया गया है। आणविक और आनुवंशिक प्रक्रिया से संबंधित कृषि अनुसंधान उपलब्ध आंकड़ों एवं परिणामों को सांख्यिकीय और कम्प्यूटेशनल तकनीकों की सहायता से विश्लेषित करने में सहायक है।

जॉब प्रस्तुतीकरण (जॉब सबमिशन): प्रस्तुत मॉड्यूल जॉब को संगठित करने और पंजीकृत सर्वरों में प्रस्तुत करने के लिए प्रयोग किया जाता है। इस तरह के मॉड्यूल में एक जॉब के रूप में कुछ सामान्य विकल्प शामिल हैं और इनपुट ब्राउज करने की क्षमता को स्थानीय रूप अच्छी तरह से सहारा देता है और उपयोगकर्ताओं के लिए लाभकारी है। उपयोगकर्ता आवश्यक जॉब प्रपत्र का चयन कर जॉब निष्पादन के लिए इस्तेमाल करने की अनुमति देता है (चित्र 3) । आवेदन से संबंधित चयन करने पर सभी अनिवार्य और वैकल्पिक विकल्प के रूप में अच्छी तरह से प्रदर्शित हो रहे हैं। फाइल प्रबंधन सेवाओं के उपयोगकर्ताओं, इनपुट फाइल /फोल्डर और लिपियों डाउनलोड, आवेदन उत्पादन फाइलें ६ फोल्डर को अपलोड फाइलों और फोल्डरों को हटाना, एक दूसरे के लिए एचपीसी संसाधन से नकल करने की अनुमति देता है (चित्र 4) ।



चित्र 3 रू जॉब निगरानी और प्रबंध

जॉब निरीक्षण (जॉब मॉनीटरिंग): उपयोगकर्ता जॉब प्रस्तुत के बाद जॉब की निगरानी एवं प्रबंध भी कर सकता है । पोर्टल प्रस्तुत जॉब की वर्तमान स्थिति के बारे में जानकारी प्रदान करता है और कोई एक निर्णय लेने के विकल्प भी प्रदान करता है। उपयोगकर्ता आसानी से विभिन्न स्थितियों के साथ इन जॉब्स का प्रबंध कर सकते हैं। जॉब प्रबंध करने के लिए जॉब आईडी, जॉब का नाम, जॉब का मालिक, जॉब की वर्तमान स्थिति, कतार, आवंटित संसाधन, वाल समय, निष्पादन नोड्स इत्यादि को पोर्टल के द्वारा नियंत्रित कर सकता है।

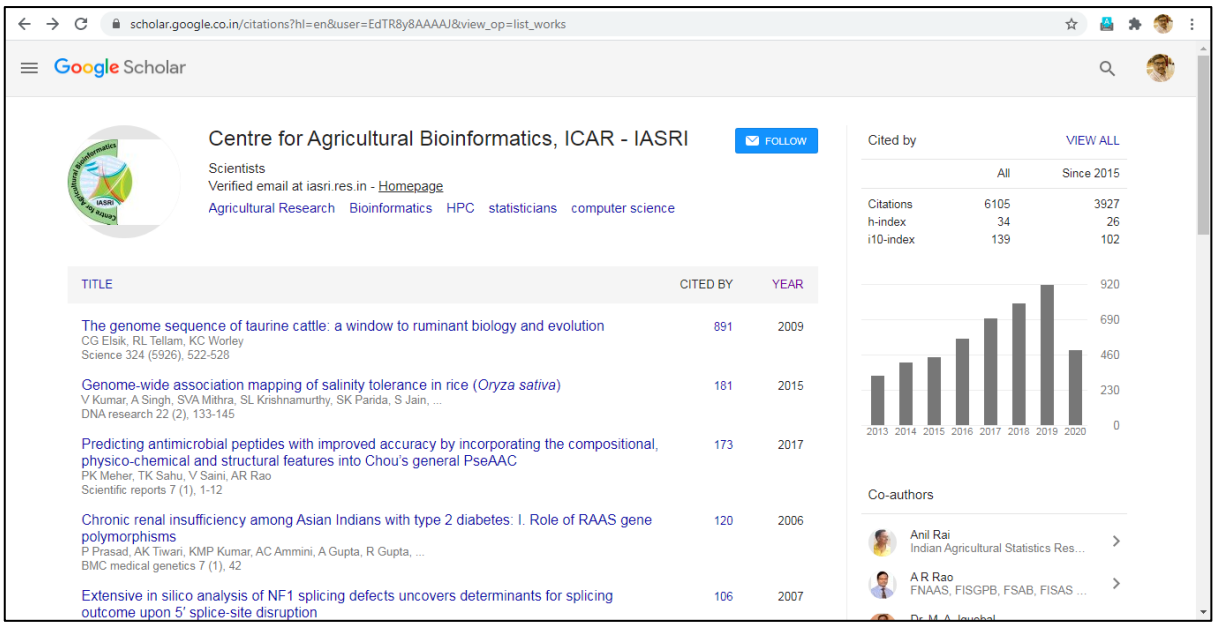


चित्र 4रू फाइल प्रबंधन

जॉब विश्लेषण (जॉब एनालिटिक्स): पोर्टल का संरक्षक (एडमिनिस्ट्रेटर) किसी भी जॉब का स्टेटस

पता लगा सकता है तथा एच पी सी में चल रही जॉब्स की प्रगति के बारे में भी जान सकता है। जॉब विश्लेषण मॉड्यूल सभी जॉब्स को एकीकृत रूप से विभिन्न ग्राफ एवं चार्ट्स के माध्यम से समझने में सहायता प्रदान करता है (चित्र ७)। इस पोर्टल की सहायता से विभिन्न पहलुओं पर एकीकृत जानकारी भी प्राप्त की जा सकती है जैसे कि (अ) कितनी जॉब्स चल रही हैं, (ब) कितनी जॉब्स समाप्त गयी हैं, (स) कितनी जॉब्स कतार में हैं, (द) प्रतिदिन कितनी जॉब्स चलती हैं (इ) कितने नोड्स फ्री और व्यस्त हैं और कई अन्य।

अशोका सुपर–कम्प्यूटर के माध्यम से कृषि जैव सूचना केंद्र ने संसथान एवं परिषद् में नए आयाम स्थापित किये हैं तथा कई उच्च कोटि के शोध पत्र भी प्रकाशित किये हैं । उच्च कोटि के प्रकाशनों की उपयोगिता को गूगल स्कॉलर के माध्यम से देखा जा सकता है (चित्र 5)।



चित्र 5 रू प्रकाशनों की उद्धरण

(स्रोतः गूगल स्कॉलर
https://scholar.google.co.in/citations?hl=en&user=EdTR8y8AAAAJ&view_op=list_works

जैव प्रौद्योगिकी बैक्टीरिया, वायरस, कवक, आदि खमीर, पशु कोशिकाओं, संयंत्र कोशिकाओं को बनाने या पौधों या जानवरों में सुधार करने के लिए या विशिष्ट उपयोगों के लिए सूक्ष्म जीवों को इंजीनियर करने के लिए अनिवार्य विषय हो गया है । विगत वर्षों में उच्च प्रदर्शन कंप्यूटिंग (एच पी सी) ने बृहद आंकड़ों को विश्लेषित करने में एक महत्तवपूर्ण योगदान दिया है और अशोका का निर्माण अपने देश के जैव वैज्ञानिकों के लिए सहायक एवं लाभकारी सिद्ध होगा । इस बड़े पैमाने पर जैविक डेटा में एन्क्रिप्टेड जैविक ज्ञान को समझने के लिए उच्च प्रदर्शन कम्प्यूटेशनल बुनियादी ढांचे में डेटा एकीकरण, फ्यूजन, खनन, कार्यप्रवाह विकास और निष्पादन, उद्गम और प्रतिनिधित्व करने के लिए इस्तेमाल किया जा सकता है। अशोका जीनोमिक डेटा संसाधनों और विभिन्न जैविक डेटाबेस के दृश्य का विश्लेषण और भंडारण के लिए उपयोगी है।

रिग्रेशन एनालिसिस तथा बेसिक स्टैटिस्टिकल टैकनीक्स

रंजीत कुमार पॉल
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

1. परिचय

समाश्रयण विश्लेषण एक सांख्यिकीय विधि जिसमे दो एवं दो से अधिक मात्रात्मक चर के बीच संबंध से एक चर का मान दूसरों चरों के आपसी संबंध से ज्ञात करते हैं । इस विधि का समान्यतः उपयोग व्यपार में, सामाजिक एवं व्यवहार विज्ञान में, जैविक विज्ञान जैसे की कृषि और मत्स्य अनुसंधान इत्यादि में करतें हैं ।

दो चरों के बीच एक कार्यात्मक संबंध को एक गणितीय सूत्र द्वारा व्यक्त करतें हैं । अगर X स्वतंत्र चर एवं Y निर्भर चर है, तो एक कार्यात्मक संबंध Y = f(X)द्वारा व्यक्त करतें हैं

X के एक विशेष मान पर, फलन f, Yके अनुरूप मान को इंगित करता है ।Xऔर Y के बीच का संबंध संबंधों की प्रकृति पर निर्भर करता है, समाश्रयण मॉडल को दो व्यापक श्रेणियां में वर्गीकृत किया जा सकता है, रेखीय समाश्रयण मॉडल एवं अरेखीय समाश्रयण मॉडल ।

2. रेखीय समाश्रयण मॉडल

हम एक बुनियादी रेखीय मॉडल पर विचार करें जिसमे केवल एक भविष्यवक्ता (predictor) चर है एवं समाश्रयण फलन रैखिक है। एक से अधिक भविष्यवक्ता चरों के साथ मॉडल निम्नानुसार दिखाया जा सकता है,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

i^{th} परीक्षण में यहाँ Y_i प्रतिक्रिया चर, β_0 एवं β_1मानकें हैं, X_i, i^{th} परीक्षण में एक भविष्यवक्ता चर के मूल्य है । ε_i एक यादृच्छिक त्रुटि है जिसका माध्य शून्य एवं विचरण σ^2है ε_i एवं ε_jअसहसंबद्ध है इसीलिए सहप्रसरण 0 है ।

2.1. समाश्रयण पैरामीटर का अर्थ

समाश्रयण मॉडल (1) में β_0और β_1मापदंडों को समाश्रयण गुणांक कहा जाता है जहाँ, β_1 समाश्रयण लाइन की ढलान है। यह X में प्रति यूनिट की बढ़ोतरी से Y के संभावना वितरणके माध्य में परिवर्तन को दर्शाता है। पैरामीटर β_0समाश्रयण लाइन Y में अन्तररोधक है ।

2.2. Least Squares विधि

β_0 और β_1समाश्रयण मापदंडों की सही अनुमान लगाने के लिए हम Least Squares विधि का प्रयोग करते हैं । Least Squares विधि मेंn squared deviationsका योग का उपयोग होता हैं । इस मापदण्ड को Qसे चिह्नित करते हैं :

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2 \qquad (2)$$

Least Squares विधि के अनुसार b_0और b_1, क्रमशःβ_0औरβ_1का अनुमानितमान है जोकिदिए गए अवलोकनों पर मापदण्ड Qकाकम से कम मान है ।

विश्लेषणात्मक दृष्टिकोण का प्रयोग से, हम समाश्रयण मॉडल (1) मेंb_0और b_1का मान जो कि किसी विशेष नमूना आँक$\overline{\text{ड़}}$ के सेट पर मापदण्ड Qका कम से कम मान से ज्ञात करते हैं तथा निम्नलिखित समीकरण द्वारा दिखाते हैं :

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 .$$

इन दोनों समीकरणों को सामान्य समीकरण कहा जाता है और b_0और b_1का मान हल किया जा सकता है:$b_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i\right) = \bar{Y} - b_1 \bar{X},$$

जहाँ $\overline{X}$ और$\overline{Y}$, X_iऔरY_iअवलोकनों का क्रमशः माध्य हैं।

2.3. युक्त/सज्जित समाश्रयण लाइन के गुण

एक बार मापदंडों का अनुमान प्राप्त हो तो , सज्जित लाइन होगा,

$$\hat{Y}_i = b_0 + b_1 X_i$$

(3)

$e_i = Y_i - \hat{Y}_i$जहाँ, e_i एकi^{th}रेसिडुअल/त्रुटिहै।

Least Squares विधि **द्वारा अनुमानित समाश्रयण लाइन (3) को सज्जित करने में निम्नलिखित गुण मिलता है,**

n

1. त्रुटियों का योग शून्य होता है, $\sum_{i=1}^{n} e_i = 0$.
2. वर्गीकृत त्रटियों का योग. e^2 न्यूनतम होता हैं।

$$\sum_{i=1} i$$

2. अवलोकित मान Y_i का योग फिट मूल्यों $\hat{Y}_i$ का योग के बराबर होती है, $\sum_{i=1} Y_i = \sum_{i=1} \hat{Y}_i$ ।
3. भारित त्रुटि का योग शून्य होता है, जहाँ X_i, i^{th} परीक्षण में भविष्यवक्ता चर के स्तर के आधार पर भारित मान है एवं $\hat{Y}_i$, i^{th} परीक्षण में प्रतिक्रिया चर के स्तर के आधार पर भारित मान है : $\sum_{i=1}^{n} X_i e_i = 0$ और $\sum_{i=1} \hat{Y}_i e_i = 0$ ।
4. समाश्रयण लाइन हमेशा बिन्दु($\bar{X}, \bar{Y}$) से गुजरता है।

2.4. पद त्रुटि के विचरण σ^2 का आकलन

Y की संभावना वितरण की परिवर्तनशीलता का एक संकेत प्राप्त करने के लिए समाश्रयण मॉडल (1) में त्रुटि पद के विचरण σ^2 का अनुमान लगाया जाने की जरूरतहोती है । साथ में, समाश्रयण फलन से संबंधित विभिन्न प्रकार का अनुमान एवं Y की भविष्यवाणी के अनुमान के लिए σ^2 का आकलनकीआवश्यकता होती है ।जिसे,

SSE = $\sum_{i=1}^{n} (Y - \hat{Y})^2 = \sum_{i=1} e^2$, द्वारा निरूपित करते हैं, जहाँSSE त्रुटि के वर्गों का योगहै।तब

त्रुटि के विचरण σ^2 का आकलन निम्नलिखित के द्वारा दिया जाता है,

$$\hat{\sigma}^2 = \frac{SSE}{n - p},$$

जहाँp मॉडल में शामिल मापदंडों की कुल संख्या है। हम इसे MSE के द्वारा निरूपित करते हैं ।

2.5. R^2 के द्वारा फिटिंग के माप निकलना

कई बार रेखीय एसोसिएशन की डिग्री को पता करने की जरूरत होती है । यहाँ हम एक वर्णनात्मक माप की प्रयोग करते है जो की कि मुख्यतः Y और X के बीच रैखिक एसोसिएशन

जहाँR^2संकल्प के गुणांक कहा जाता है,$0 \le R^2 \le 1$. इसका मान 1 के करीब हो तो यहअधिक से अधिकYऔरX के बीच रैखिक एसोसिएशन की डिग्री देता है ।

2.6. निदानिकी एवं उपचारी उपाय

जब हम एक समाश्रयण मॉडल लेते हैं, तो हम आम तौर पर यह नहीं होताहैकि अग्रिम में निश्चित हो कि लिया गया मॉडल किसी विशेष अनुप्रयोग के लिए उपयुक्त है, किसी भी मॉडल कीएक या कई गुण जैसे की समाश्रयण फलन में रैखिकता हो सकता है या त्रुटि के संदर्भ में प्रसामान्यथता (normality) होना, विशेष रूप के डेटा के लिए उपयुक्त नहीं हो सकता है । इस भाग में हम एक मॉडल के औचित्य के अध्ययन के लिए कुछ सरल ग्राफिक तरीकों, साथ ही कुछ सुधारात्मक उपाय जब डेटा समाश्रयण मॉडल की शर्तों के अनुसार सहायक नहीं हो पर चर्चा करेंगे ।

2.7. मॉडल से प्रस्थापन का अध्ययन

हम सामान्य त्रुटियों के साथ रेखीय समाश्रयण मॉडल से प्रस्थापन के छह महत्वपूर्ण प्रकार निम्नलिखित पर विचार करेंगेरू (i) समाश्रयण फलन के रैखिकता ।

(ii) त्रुटि विचरण की स्थिरता ।

(iii) त्रुटि पदों की स्वतंत्रता।

(iv) एक या कुछ गैर अवलोकन की उपस्थिति ।

(v) त्रुटि पदों की सामान्य वितरण (normal distribution)।

(vi) एक या कई महत्वपूर्ण भविष्यवक्ता चर मॉडल से मिटाया गया हो ।

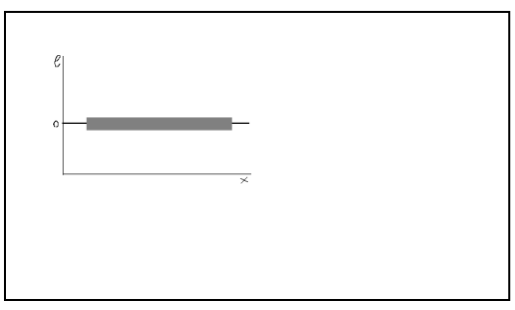
(vii) multicollinearityकी उपस्थिति।

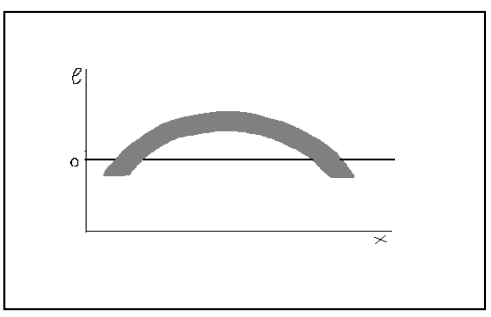
2.8. मॉडल से प्रस्थापन के लिए ग्राफिकल टेस्ट

समाश्रयण मॉडल की गैर रैखिकता

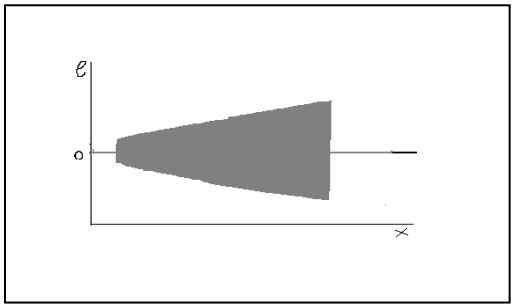
आकृति 1(a) एक रेखीय समाश्रयण मॉडल का उचित प्रोटोटाइप स्थिति को दिखता है । यहाँresidualएक क्षैतिज बैंडमें है जो 0 के आसपास केंद्रित है और कोई व्यवस्थित प्रवृत्तियों के लिए सकारात्मक और नकारात्मक को प्रदर्शित नहीं करता है ।

आकृति 1(b) रेखीय समाश्रयण मॉडल से प्रस्थापन का एक प्रोटोटाइप स्थिति दिखता है जो कि एक वक्रीय समाश्रयण फलन के लिए की आवश्यकता को इंगित करता है । यहाँ residual सकारात्मक और नकारात्मक के बीच एक व्यवस्थित तरीके में बदलते हैं ।



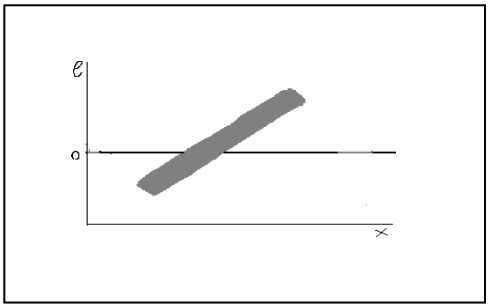


आकृति1(a)



आकृति1(c)

आकृति1(b)



आकृति1(d)

त्रुटि विचरण की गैर स्थिरता

आकृति 1(a) में प्रोटोटाइप प्लाट दिखता है की त्रुटि पद विचरण स्थिर है, एवंआकृति1(c) में प्रोटोटाइप प्लाट दिखता है की त्रुटि पद विचरण X के साथ बढ़ती जाती है जो की"Megaphone"प्रकार दीखता है ।

Outliersकी उपस्थिति

Outliers चरम अवलोकन है । अवशिष्ट(residual)outliers,residualप्लाट Xया$\hat{Y}$ के खिलाफ से पहचाना जा सकता है।

त्रुटिपदों के गैर स्वतंत्रता

आकृति 1(d) में एक प्रोटोटाइप अवशिष्ट प्लाट समय संबंधित प्रवृत्ति असर दिखाता है, जो एक रेखीय समय संबंधित प्रवृत्ति प्रभाव का चित्रण है।

त्रुटि पदों के गैरNormality

त्रुटि पदों के Normalityका अध्ययन विभिन्न ग्राफिक तरीकों से residuals परीक्षण से किया जा सकता है । जैसे की आवृत्तियों की तुलना एवं सामान्य संभावना प्लाट।

महत्वपूर्ण कारक चर की अनुपस्थिति

अतिरिक्त भविष्यवक्ता चर के खिलाफ residualplot से यह पता करते है की अतिरिक्त भविष्यवक्ता चर के विभिन्न स्तर के साथ व्यवस्थित ढंग से भिन्न residual trend है या नहीं ।

2.9. मॉडल से प्रस्थापन के लिए सांख्यिकीय परीक्षण

Randomness के लिए टेस्ट

Durbin-Watson test:अगर ρ ,autocorrelation गुणांक है और $e_t = Y_t - \hat{Y}_t$ तो

$H_0 : \rho = 0$

$H_1 : \rho > 0$

n

$$D = \frac{\sum_{t=2}^{} (e_t - e_{t-1})^2}{n\sum_{t=1} e_t^2}$$

निष्कषः D के छोटे मान से यह पता चलता है कि$\rho>0$।

Normalityके लिए परीक्षण

Normalityके लिए Correlationपरीक्षण: Kolmogorov-Smirnov testऔरAnderson-Darling Test.

Tests for Constancy of Error Variance: Modified Levene Test औरWhite Test

Outlyingअवलोकनों के लिए परीक्षण :

(i) **Elements of Hat Matrix:** $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, जहाँ**X**व्याख्यात्मक चर के लिए मैट्रिक्स है, जो

बड़ मानों के आँकड़ बिंदु,outliers को दर्शाते हैं ।

(ii) ***WSSD**i*: x-spaceमें किसी दूरदराज के बिंदु का पता लगाने के लिए*WSSD*$_i$ का उपयोग करते हैं ।

(iii) ***Cook's D**i:* Cook's D$_i$ $\hat{y}$ में परिवर्तन को मापता है जब ithअवलोकन मापदंडों के आकलन में इस्तेमाल नहीं किया गया हो ।

(iv) ***DFFITS**i :* *DFFIT*का उपयोग $\left(\hat{y} - \hat{y}_{(i)}\right)$के ithघटक में अंतर मापने में होता है ।

(v) $\boldsymbol{DFBETAS}_{j(i)}$: व्यक्तिगत समाश्रयण गुणांक के लिए प्रभावशाली अवलोकनों

$DFBETAS_{j(i)}, j = 1,2,...,p+1$, द्वारा पहचाने जाते हैं ।

(vi) $\boldsymbol{COVRATIO}_i$ **:**अनुमानित समाश्रयण गुणांक के विचरण–सहप्रसरण मैट्रिक्स पर*i*thअवलोकन के प्रभाव को दो विचरण–सहप्रसरणmatricesके निर्धारकों के अनुपात से मापा जाता है। इस प्रकार, COVRATIOसमाश्रयण गुणांक के अनुमानों की शुद्धता पर*i*thअवलोकन के प्रभाव को दर्शाता है।

(vii) **FVARATIO** $_i$ **:** जब एक अवलोकन हटाया जाता है तो $\hat{y}_i$ के विचरण में परिवर्तन का पताFVARATIO$_i$ से लगाते है ।

2.10. Multicollinearity

Multicollinearityकी समस्याओं को निम्नलिखित तरीकों से दूर कर सकतें हैं **-** ***अतिरिक्त डेटा का संग्रह,***Model respecificationऔरRidge Regression.

उदाहरण:

तालिका 1

Case	X_{11}	X_{21}	X_{31}	Y_i
1	12.980	0.317	9.998	57.702
2	14.295	2.028	6.776	59.296
3	15.531	5.305	2.947	56.166
4	15.133	4.738	4.201	55.767
5	15.342	7.038	2.053	51.722
6	17.149	5.982	-0.055	60.446
7	15.462	2.737	4.657	60.715
8	12.801	10.663	3.048	37.447
9	17.039	5.132	0.257	60.974
10	13.172	2.039	8.738	55.270
11	16.125	2.271	2.101	59.289
12	14.340	4.077	5.545	54.027
13	12.923	2.643	9.331	53.199
14	14.231	10.401	1.041	41.896
15	15.222	1.220	6.149	63.264
16	15.740	10.612	-1.691	45.798
17	14.958	4.815	4.111	58.699
18	14.125	3.153	8.453	50.086

19	16.391	9.698	-1.714	48.890
20	16.452	3.912	2.145	62.213
21	13.535	7.625	3.851	45.625
22	14.199	4.474	5.112	53.923
23	15.837	5.753	2.087	55.799
24	16.565	8.546	8.974	56.741
25	13.322	8.589	4.011	43.145
26	15.949	8.290	-0.248	50.706

तलिका 2:प्रभावशाली प्रेक्षणों के संकेतक

Case	r_i	t_i	$t_i^*=s.t/s_i$	h_{ii}	D_i	$WSSD_i$
1	0.460	0.289	0.281	0.215	0.005	39*
2	1.253	0.732	0.724	0.093	0.013	12
3	0.377	0.215	0.210	0.048	0.001	1
4	0.044	0.025	0.026	0.042	0.000	1
5	-0.256	-0.146	-0.141	0.053	0.000	3
6	1.010	0.611	0.602	0.155	0.017	20
7	0.389	0.226	0.221	0.081	0.001	7
8	0.132	0.088	0.086	0.301	0.001	41
9	0.432	0.262	0.256	0.155	0.003	18
10	0.589	0.355	0.347	0.147	0.005	23
11	-3.302	-2.021	-2.193	0.173	0.214	14
12	-0.406	-0.232	-0.226	0.053	0.001	3
13	0.194	0.118	0.117	0.163	0.001	24

14	-0.268	-0.164	-0.161	0.175	0.001	23
15	0.802	0.476	0.469	0.122	0.007	15
16	-0.482	-0.295	-0.289	0.177	0.005	26
17	3.756	2.134	2.343	0.041	0.048	0
18	-6.072	-3.589	-5.436	0.114	0.412	8
19	-1.198	-0.727	-0.719	0.160	0.025	24
20	1.126	0.666	0.658	0.114	0.014	11
21	0.449	0.266	0.259	0.119	0.003	12
22	0.791	0.453	0.444	0.055	0.003	3
23	-0.060	-0.035	-0.032	0.059	0.000	3
24	0.574	1.181	1.188	0.927	4.409	19
25	0.268	0.163	0.158	0.159	0.001	19
26	-0.606	-0.356	-0.350	0.101	0.004	11

तलिका 3:प्रभावशाली प्रेक्षणों के संकेतक

Case	Cov Ratio	Dffits	Intercept	X1	X2	X3
				DFBETAS		
1	1.512	0.148	0.056	-0.053	-0.006	0.006
2	1.203	0.232	0.062	-0.042	-0.042	-0.050
3	1.254	0.047	-0.005	0.010	-0.008	-0.007
4	1.257	0.005	0.000	0.000	-0.001	0.000
5	1.267	-0.033	-0.001	-0.001	-0.006	0.006
6	1.331	0.258	-0.095	0.132	-0.042	-0.050
7	1.299	0.068	-0.005	0.015	-0.036	-0.005

8	1.721	0.057	0.027	-0.034	0.026	-0.006
9	1.408	0.109	-0.030	0.048	-0.035	-0.031
10	1.380	0.144	0.058	-0.058	-0.041	0.016
11	0.639	-1.004	-0.154	-0.045	0.776	0.525
12	1.260	-0.054	-0.017	0.014	0.014	0.000
13	1.435	0.051	0.017	-0.19	-0.004	0.013
14	1.452	-0.074	-0.026	0.031	-0.35	0.015
15	1.315	0.175	-0.008	0.033	-0.105	0.002
16	1.441	-0.134	-0.014	0.014	-0.044	0.047
17	0.496	0.482	0.061	-0.17	-0.107	-0.046
18	0.410	-1.945	0.362	-0.308	-0.220	-1.177
19	1.301	-0.341	0.031	-0.045	-0.080	0.094
20	1.252	0.236	-0.055	0.097	-0.105	-0.051
21	1.350	0.095	0.054	-0.061	0.024	-0.018
22	1.228	0.108	0.052	-0.048	-0.028	-0.020
23	1.279	-0.008	0.001	-0.002	0.001	0.002
24	12.715	4.230	-3.642	3.276	3.180	3.934
25	1.426	0.069	0.031	-0.039	0.029	-0.003
26	1.309	-0.117	0.000	-0.007	-0.016	0.043

तलिका 4:समाश्रयण गुणांक और सारांश आँकड़े

Description	b_0	b_1	b_2	b_3	s	R^2	Max VIF	Min e.v.	Max R_i^2
All Data (n=26)	8.11	3.56	-1.63	0.34	1.80	0.94	2.82	0.210	0.65
Delete (11, 17, 18)	7.17	3.66	-1.79	0.40	0.51	0.99	2.85	0.210	0.65

Delete (24)	30.91	2.39	-2.14	-0.36	1.78	0.94	30.64	0.017	0.97
Delete (11, 17, 18, 24)	24.27	2.79	-2.11	-0.16	0.50	0.99	171.90	0.003	0.99
Ridge k=0.05 (n=22)	14.28	3.22	-1.73	0.25	0.66	0.99	10.20	0.053	0.90
Delete X3 (n=22)	19.50	3.03	-2.00		0.49	0.99	1.02	0.863	0.02

कुछ चयनित संदर्भ

वेल्स्ली, डी.ए., कुह, ई एवं वेल्श, आर.ई. (2004): "समाश्रयण निदान – प्रभावशाली डेटा एवं collinearity, के स्रोतों की पहचान", न्यू यॉर्क.: विले लिमिटेड ।

बार्नेट, वी और लुईस, टी (1984): "सांख्यिकीय डेटा में outliers", न्यू यॉर्कः विले लिमिटेड ।

चटर्जी, एस और मूल्य, बी (1977): "उदाहरण के द्वारा समाश्रयण विश्लेषण", न्यू यॉर्कः जॉन विले एंड संस ।

ड्रेपर, एन.आर. एवं स्मिथ, एच (1998): "एप्लाइड समाश्रयण विश्लेषण", न्यू यॉर्कः विले पूर्वी लिमिटेड ।

क्लेंबौम, डी.जी. ,वं कुप्पेर, एल.एल. (1978): "एप्लाइड समाश्रयण विश्लेषण और अन्य बहुभिन्नरूपी चर तकनीक", मैसाचुसेट्सः देक्स्वरी प्रेस ।

मांटगोमेरी, डी.सी., पेक, एवं भिनिंग, जी (2003): "रेखीय समाश्रयण विश्लेषण का परिचय ", तीसरा संस्करण, न्यू यॉर्कः जॉन विले एंड संस ।

जेनेटिक पैरामीटर एस्टिमेशन

डॉ. अमृत कुमार पॉल

भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

"पशु डेरी में असंतुलन आँकड़ों के लिए स्टेएबिलिटी की वंशागतित्व के अनुमान की विभिन्न प्रक्रियाओं की अनुभाविक तुलना"

(अमृत कुमार पॉल,)

सारांश

स्टेएबिलिटी (Stayability)जोकि पशु प्रजनन में एक द्विगुण विशेषक है। वह जननिक विश्लेषण के माध्यम से आवश्यक है। **स्टेएबिलिटी** की सही माप के लिए उत्पादन तथा उत्पन्न करने वाले विशेषक के लिए समायोजित करने की आवश्यकता है। **स्टेएबिलिटी** की वंशागतित्व का आकलन हेतु उत्पादन के लिए समायोजित पशुओं के झुण्ड जीवन का बदलाव द्विगुण विशेषक में प्रारम्भिक प्रायिकता का प्रयोग करके किया गया है। प्रारम्भिक गुणो की व्यापक कार्यविधि फलकोनर और मैकी [3] द्वारा रूपरेखित की गयी है। पशुओं के झुण्ड जीवन के समायोजन को समाविष्ट करने के लिए बीटा–द्विपद की प्रक्रिया को संशोधित किया गया डैमस्टर लरनर[2] की विधि का भी प्रयोग **स्टेएबिलिटी** की वंशागतित्व का अनुमान करने के लिए किया गया तथा साथ–साथ इसकी तुलना प्रयोगजन्य (impirically) के आधार पर बीटा–द्विपद विधि से की गयी है। इन सभी तुलनाओं के लिए असंतुलित आँकड़ों की स्थिति को ध्यान में रखा गया है। यह देखा गया है कि उत्पादन के संदर्भ में छोटी सी समायोजिता का **स्टेएबिलिटी** की वंशागतित्व के अनुमानो पर बहुत बड़ा प्रभाव है। असंतुलिता के होने पर अनुमानो के बड़ी मानक त्रुटि होने की संभावना होती है। इसके आगे आपेक्षित रूट माध्य वर्ग (Relative root mean square) भी प्राप्त किया गया और यह पाया गया कि आनुमानो की परिशुद्धता व यथार्थता उत्पादन के समायोजन से प्रभावित थे। आनुमानो की बीटा–द्विपद विधि के परिणामों ने कुछ एक रूपता दिखायी जबकि बाकी प्रक्रियाओं में उत्पादन के आँकड़ों के समायोजन के लिए कोई विशिष्ट प्रवृति नहीं देखी गयी ।

परिचय

पशु प्रजनन में ज्यादातर द्विगुण विशेषक(Binary Traits) उत्पादन दक्षता(efficiency) के महत्वपूर्ण निर्धारक(Determinants) होते हैं और मूलभूत कारकों के सूचक हैं जिनकी माप मुश्किल या मंहगी होती हैं। लक्षणों मामले में जिनका दृश्यप्रारूप(phenotype) सामान्य संस्थापक(classical) विधियों द्वारा अभिव्यक्त किया जाता है। यह सीधे प्रयुक्त नहीं किये जाते। वह लक्षण जिनकी वंशागति बहु–उपादानीय (multifactorial) होती है लेकिन जिनमें किसी भी प्रकार नहीं या सभी प्रकार की दृश्य प्रारूप अभिव्यक्त होते हैं। वह प्रारम्भिक लक्षण कहलाते हैं। एक ऐसा ही लक्षण जो पशु डेरी में प्रारम्भिक की तरह वर्गीकृत है वह गाय के झुण्ड में **स्टेएबिलिटी** है। अगर गाय रखी जाती है तो यह माना जाता है कि वह पशुओ के झुण्ड में रहेगी अन्यथा अलग कर दी जायेगी। या तो किसी प्रकार नहीं

या सभी प्रकार की विशेषक **स्टेएबिलिटी** की वंशागतित्व के अनुमान को सरल बनाने के लिए प्रारम्भिक मॉडल मान लिया जाता है। प्रारम्भिक लक्षणों की वंशागतित्व के अनुमान की अनेक विधियाँ है लेकिन सभी विधियाँ असंतुलित आँकड़े समुच्चय के लिए सीधे तरीके से प्रयुक्त नहीं होती है। इस दृष्टिकोण के लिए डैम्पस्टर लरनर की दोनों बीटा–द्विपद विधियों की असंतुलित आँकडों के मामले में तुलनात्मक निष्पादन के अध्ययन का प्रयास किया गया है।

आँकडें मॉडल

इससे पहले की दो प्रक्रियाओं की संकल्पनाओं पर विचार विमर्श किया जाये, **स्टेएबिलिटी** के आँकड़े इस प्रकार है। दी हुयी जनसंख्या में प्रक्रिया को मानकीकृत गोसियन विचर(Z) के द्वारा समझाया गया है। जिसका माध्य शून्य है तथा प्रसरण एक है। जब भी z एक प्रारम्भिक संख्या से ज्यादा हो जाता है। तब उसे z' मानेंगे, जोकि ज्ञात है। तब एक बाध्य आवलोकन लक्षण (δ)अभिव्यक्त किया जाता है। यह लक्षण का मान उपस्थिति पर एक तथा अनुपस्थिति पर शून्य है।

अनुपालनीय विचर (z) के लिए रेखीय मॉडल

$$Z_{ijk} = \mu + S_i + e_{ijk} \qquad \ldots\ldots(1)$$

जहाँ z_{ijk} j^{th} ब्लॉक के i^{th} परिवार के k^{th} व्यक्तिगत पर अवलोकन है।

μ सम्पूर्ण माध्य है।

S_i i^{th} परिवार प्रभाव है।

e_{ijk} अवशिष्ट प्रभाव है जोकि पिलोट ब्लॉक और त्रुटि प्रभाव से गठित है।

$S_i \sim N(0, \sigma_f^2$ ढ vkSj $e_{ijk} \sim N(0, \sigma_e^2)$

वास्तविक विचर(intrinsic variable)का बाह्य स्कैल(outward scale) पर द्विपद विशेषक (δ) में रूपान्तरण इस प्रकार किया गया।

$$\delta_{ijk} = 1 \quad for \quad Z_{ijk} \leq Z' or\ \Phi(Z_{ijk}) \leq P$$

$$= 0 \quad for \quad Z_{ijk} > Z' or\ \Phi(Z_{ijk}) > P$$

जहाँ φ एक सामान्य वितरण संचीय प्रायिकता फलन को दर्शाता है तथा p डाइकोटोमस लक्षण(δ) की निरीक्षण की जनसंख्या प्रायिकता बताता है।

डैमपस्टर और लरनर विधि

डैमपस्टर और लरनर [2] ने द्विगुण विशेषक के लिए व्यक्तिगत संकीर्ण संवेदी वंशागतित्व का अनुमान दिया जोकि h_{DL}^2 द्वारा निर्दिष्ट किया गया है। जिसको ज्ञायानोला [4] ने अधिक सामान्य समाधान(Solution) के विशेष मामले की तरह दिखाया है।

जैसे

$$\hat{h}^2_{DL} = 4\hat{\sigma}^2_f(\delta) \times [\phi(Z')]^{-2} \quad \text{...(2)}$$

जहाँ φ द्विगुण स्कैल $[Z' = \phi^{-1}(p)]$ पर व्यंजक के लिए प्रारम्भिक पर मूल्यांकित किया गया जो गोसियन प्रायिकता घनत्व फलन को दर्शाता है तथा $\sigma^2_i(f)$ परिवार प्रसरण घटक का अनुमान है जोकि द्विगुण विशेषक पर प्रयोग किया गया प्रसरण विधि (ANOVA) के विश्लेषण से प्राप्त किया गया है।

बीटा–द्विपद मॉडल प्रस्ताव

निम्नलिखित मैगनूसेन और बरेमर[5] बीटा–प्राचल के तीन समुच्चयः एक दृश्य प्रारूप (phenotypic) परिवार प्रायिकताओं के लिए, एक परिवार प्रायिकताओं के लिए तथा एक संयोजी (additive) जननिक प्रायिकता के लिए δ_{ijk} एक द्विगुण–विशेषक आँकड़ो के मॉडल पर आधारित, वंशागतित्व अनुमानो पर आधारित प्राप्त करने के लिए कल्पित है।

$$p_{ijk} = p + p_i + p_{ijk} \quad \text{...(3)}$$

जहाँ p_{ijk} i परिवार के j^{th} ब्लॉक में k^{th} व्यक्तिगत पर द्विगुण विशेषक ,δद्ध की निरीक्षणता की प्रायिकता है, p सम्पूर्ण जनसंख्या प्रायिकता(स्थिर प्रभाव) है और शेष p परिवार प्रभाव तथा अवशिष्ट के क्रमशः रेन्डम योगदान है। इस मॉडल दृश्य प्रारूप [(pf), परिवार (f) और संयोजी जननिक(a)] से तीनों प्रसरण घटक $\sigma_{pf}^2(d)$, $\sigma_f^2(\delta)$ और $\sigma_a^2(\delta)$ द्विगुण आँकड़ों (δ_{ijk}) पर किये गये प्रसरण एक तरफा विश्लेषण द्वारा प्राप्त किया जाता है।

दृश्य प्रारूप परिवार प्रायिकतायें ख्च्,च्/द्विप, एक बीटा–वितरण का अनुसरण करता है।

$$p_{(pf)i} = \sum_{jk} \frac{P_{ijk}}{n_i..} \quad \text{lkFk} \quad p_{(pf)i} \sim Beta\ (\alpha_{pf}, \beta_{pf}) \quad \text{...(4)}$$

जहाँ d_{qvv} परिवार i में की गयी अवलोकनों की संख्या है उसी तरह परिवार प्रायिकतायें सम्पूर्ण माध्य तथा संयोजित (aaditive) परिवार प्रभाव एक योग (sum) द्वारा परिभाषित (defined) की जाती है

$$p_{(f)i} = p + p_i \quad \text{with } p_{(f)i} \sim Beta\ (\alpha_f, \beta_f) \quad \text{.....(5)}$$

यह मानते हुये कि परिवार में हॉफ–सिब है, निम्नलिखित संकलपना–संबंधी (conceptual) मॉडल संयोजी जननिक परिवार प्रायिकताओं $[p_{(a)i}]$ के लिए प्रयोग होता है।

$$p_{(a)i} = p + 0.5p_i \text{ with } p_{(a)i} \sim Beta\ (\alpha_a, \beta_a) \qquad \alpha_a, \beta_a > 0 \quad \text{....(6)}$$

उपरोक्त प्रायिकताओं के नमूना अनुमानो आँकड़ों से इस प्रकार प्राप्त किये गये हैं।

$$\hat{p}_{(pf)i} = \sum_{jK} \frac{\delta_{ijk}}{n_i..} = \hat{p}_{(f)i} \quad \text{.....(7)}$$

$$\bar{p} = \sum_i \frac{\hat{p}_{(pf)i}}{n_{fam}} \quad \text{.....(8)}$$

निम्नलिखित जोनशन एण्ड कौनज[5], तीन समुच्चयों के प्राचलो का अनुमान निम्नलिखित तरीकों से प्राप्त किया जा सकता है बीटा–वितरण का परिवार निम्नलिखित तरह के प्रायिकता घनत्व फलन के सभी वितरणों से प्रकृतिस्थ (composed) है।

$$P_{y(Y)} = \frac{1}{B(\alpha, \beta)} \frac{(Y - a)^{\alpha-1}(b - y)^{\beta-1}}{(b - a)^{\alpha+\beta-1}} \quad ;a \le y \le b), \alpha, \beta > 0 \qquad \text{..... (9)}$$

वितरण (4.7)में सभी चारों प्राचलो का अनुमान नमूनों की बराबरी करके तथा पहले चार आघूर्ण (moment) की जनसंख्या के मान से प्राप्त किया जा सकता है।

यदि a और b के मान ज्ञात है तब पहला और दूसरा आघूर्ण इस प्रकार दिये जाते हैं।

$$\mu_1' = \frac{a + (b - a)\alpha}{\alpha + \beta} \qquad \text{.....(10)}$$

$$\mu_2 = (b - a)^2 \alpha\beta(\alpha + \beta)^{-2}(\alpha + \beta + 1)^{-1} \qquad \text{....(11)}$$

जहाँ

$$\frac{\mu_1'-a}{b-a} = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \frac{\mu_2}{(b-a)^2} = \frac{\alpha}{\alpha+\beta}\left(1-\frac{\alpha}{\alpha+\beta}\right)\frac{1}{\alpha+\beta+1} \qquad \text{.....(12)}$$

तो अब (Thus)

$$\alpha + \beta = \frac{u_1'-a}{b - a} \frac{\left(1 - \frac{u_1'-a}{b - a}\right)}{\left(\frac{\mu_2}{(b - a)^2}\right)} - 1 \qquad \text{…(13)}$$

$$\alpha = \left(\frac{u_1'-a}{b - a}\right)^2 \left(1 - \frac{u_1'-a}{b - a}\right) \left(\frac{\mu_2}{(b - a)^2}\right)^{-1} - \frac{u_1'-a}{b - a} \qquad \text{…(14)}$$

a को 'kwU; और b को एक लेकर उपरोक्त समीकरण घटकर कुछ इस प्रकार है

$$\alpha = \mu_1^{2}, \frac{(1 - \mu_1')}{\mu_2} - \mu_1' \qquad \text{…(15)}$$

$$\alpha + \beta = \mu_1' \frac{(1 - \mu_1')}{\mu_2} - 1 \qquad \text{…(16)}$$

अब समीकरण (15) और (16) को हल करके तथा मान μ_1' त्र $\bar{P}$ और μ_2 त्र $\sigma_{ज}^2$ को रखते हुये तीन समुच्चो का अनुमान सामान्य तरीके में इस प्रकार लिखा जा सकता है

$$\hat{a}_t = \bar{P}^2 \times \frac{(1-\bar{P})}{\hat{\sigma}_t^2(\delta)} - \bar{P} \qquad \ldots(17)$$

$$\hat{\beta}_t = \frac{(1-\bar{P})}{\bar{P}} \hat{\alpha}_t \qquad \ldots(18)$$

जहाँ पर सब्सक्रिप्ट t प्राचल [t=f (परिवार), pf (दृश्य प्रारूप परिवार माध्य)] या परिवार संयोजी जननिक)] के प्रकार को दर्शाता है और $\hat{\sigma}_t^2(\delta)$ अनुकरणीत द्विगुण–विशेषक (δ) के प्रसरण के विश्लेषण से अनुमानित समरूप प्रसरण घटक है।

सरले इत्यादि का अनुसरण करते हुये प्रतिबंधित माध्य परिवार प्रायिकता चुनी हुयी जनसंख्या में कुछ इस प्रकार अपेक्षा की जाती है।

$$\bar{p}_{t/\delta} = \frac{\delta + \hat{\alpha}_t}{1 + \hat{\alpha}_t + \hat{\beta}_t} \qquad \text{---(19)}$$

जहाँ t=(f,pf,a) सोच–विचार के अन्तर्गत प्रभाव को दर्शाता है माध्य में चयन के कारण बदलाव $(p_{t/\delta} - \bar{p})$ को चयन पर अनुक्रिया (प्रतिक्रिया) माना जा सकता है। जिससे चयन के अन्तर्गत विशेषक की अपेक्षा उपलब्धी वंशागतित्व का अनुमान किया जा सकता है।

चयन प्रतिक्रिया का अनुमान

$$\Phi^{-1}(p_{t/\delta}) - \Phi^{-1}(\bar{p}) \qquad \ldots(20)$$

उपलब्ध (Realized) व्यक्तिगत संकीर्ण संवेदी वंशागतित्व का बीटा–द्विपद अनुमान इस प्रकार है

$$h^2 = \frac{\text{चयन पर आपेक्षित प्रतिक्रिया}}{\text{दृश्य प्रारूपित चयन विभेदन}} \qquad \ldots.(21)$$

बीटा वितरण प्राचल संयोजी परिवार प्रसरण तथा दृश्य प्रारूप परिवार प्रसरण का अनुपात लेते हुये परिवार माध्य वंशागतित्व को संगणक करने के लिये प्रयोग किया जा सकता है।

$$h^2_{f(beta)} = \frac{\hat{\alpha}_f \times \hat{\beta}_f \times \left(\hat{\alpha}_{pf} + \hat{\beta}_{pf}\right)^2 \times \left(\hat{\alpha}_{pf} + \hat{\beta}_{pf} + 1\right)}{\hat{\alpha}_{pf} \times \hat{\beta}_{pf} \times \left(\hat{\alpha}_f + \hat{\beta}_f\right)^2 \times \left(\hat{\alpha}_f + \hat{\beta}_f + 1\right)} \qquad \ldots(22)$$

प्राचल a और b के साथ एक बीटा वितरण का प्रसरण है।

$$\frac{ab}{(a+b+1)(a+b)^2}$$

बीटा–द्विपद मॉडल में जड़ित परिवार माध्य वंशागतित्व का एक वैकल्पिक फार्मूला परिवार माध्य प्रायिकता$(\bar{p}_{f/\delta} - p)$ में उपलब्ध चयन प्रतिक्रिया तथा दृश्य प्रारूप परिवार माध्य लैवल $(\bar{p}_{f/\delta} - p)$ चयन प्रतिक्रिया के अनुपात से प्राप्त किया जाता है। संचीय वितरण फलन के द्धारा इन प्रतिक्रियों के अनुपात को वास्तविक विचर (z) के स्केल (scale) पर बदला जाता है जोकि उपलब्ध परिवार माध्य वंशागतित्व के अनुमान की उपज करता है।

$$h^2_{f(\Delta P/beta)} = \frac{\Phi^{-1}\left(\bar{p}_{f/\delta}=1\right)-\Phi^{-1}\left(\bar{p}\right)}{\Phi^{-1}\left(\bar{p}_{pf/\delta}=1\right)-\Phi^{-1}\left(\bar{p}\right)} \quad \text{...23}$$

सहायक विशेषकों के लिए स्टेएबिलिटी का समायोजन

लक्षणों की तरह **स्टेएबिलिटी** उत्पादन जैसे सहायक लक्षण तथा अन्य प्रकार के लक्षणों द्वारा सार्थक तरीके से प्रभावित होते हैं इसलिए **स्टेएबिलिटी** की वंशागति की सही तस्वीर पाने के लिए सहायक लक्षणों के प्रभाव को निरसन(eliminate) करने की सलाह दी जाती है। उदाहरण के तौर पर पशु डेरी के झुण्ड जीवन उत्तरजीविता(survival) तथा उत्पादन विशेषक के रूप में होती है, जोकि निम्नलिखित समीकरण द्धारा निर्देशित है।

$$P_{HL} = m_Y P_Y + m_S P_S \quad \text{.....(24)}$$

जहाँ p_{HL}, p_y, p_s क्रमशः झुण्ड जीवन के दृश्य प्रारूप मान, उत्पादन तथा उत्तरजीविता है। क्रमशः p_y और p_s पर p_{HL} के मानवीकृत आंशिक समाश्रयण गुणांक (standardized partial regression coefficient) m_Y और m_S हैं। उत्पादन के लिए समायोजित झुण्ड जीवन के एक नये दृश्य प्रारूप विचर कुछ इस प्रकार आसानी से प्राप्त किये जा सकते हैं।

$$P_{HL/Y} = P_{HL} - r_{Y,HL} P_Y \quad \text{.....(25)}$$

$$= m_S(P_S - r_P P_Y)$$

अब $p_{HL/Y}$ को असली विचर लेते हुए, नये विचर को दी गयी सफलता की प्रायिकता के लिए रुड़न के विभिन्न बिन्दुओं की मदद से द्विपद विचर में बदला जाता है, उत्पादन के लिए समायोजित पशुओं के झुण्ड जीवन की वंशागतित्व का अनुमान आसानी से प्राप्त किया जा सकता है। अब समायोजित लक्षणों की वंशागतित्व का अनुमान वंशागति की सही तस्वीर प्रदर्शित करेगा जबकि पशुओं के झुण्ड जीवन की वास्तविक मान सहायक लक्षण उत्पादन द्धारा सार्थक तरीकों से प्रभावित हो सकते हैं।

आपेक्षित मूल माध्य वर्ग त्रुटि

विभिन्न विधियों की तुलना उसके परिशुद्वता की कुछ मापदण्डों के आधार पर की गयी है क्योंकि सभी अनुमान अनभिन्नत नहीं हैं। इसलिए प्रसरण के अनुमान सही अंदाजा नहीं दे सकते। अभिनत तथा कुछ परिशुद्धता के मापों के मान जानने के लिए, एक माप जोकि आपेक्षित रूट माध्य वर्ग त्रुटि के द्धारा कुछ इस प्रकार परिभाषित किया जाता है।

$$RMSE\,\% = \frac{\left[E\left(estimate - 'true\ value'\right)^2\right]^{0.5}}{'true\ value'} \times 100 \quad \text{.....(26)}$$

असंतुलितता की मात्रा

असंतुलितता की मात्रा को इस प्रकार परिभाषित कर सकते हैं

$\Delta=N(n-\lambda)$ tgWk $n= N/S$, $\sum_{i=1}^{S} n_i = N$

$$\lambda = \frac{1}{S-1}\left[\sum_i n_i - \frac{\sum_i n_i^2}{N}\right]$$

यॅहा S = जनको या साड़ की संख्या

n_i = जनक (साड़) की पुत्री की संख्या

N = सम्पूर्ण पुत्रियों की संख्या

रॉा आँकड़ो पर वंशागतित्व का अनुमान

मोन्टे कारलो अनुकरण द्वारा जनित आंकड़े ऑफ–सिब मॉडल अनुसरण करता है।

$$Z_{ijk} = \mu + S_i + e_{ijk}$$

सही वंशागतित्व या रॉ आँकड़ों पर आधारित वंशागतित्व वह वंशागतित्व है जो कि द्विपद आँकड़ो या प्रारम्भिक लक्षणों को असली ऑफ–सिब अनुकरणीत आँकड़ों का प्रयोग करके संगणित की गयी है या की जाती हैं।

व्यक्तिगत संकीर्ण संवेदिये वंशागतित्व

$$\hat{h}^2_{(Z)} = \frac{4\hat{\sigma}^2_f(z)}{\hat{\sigma}^2_f(z)+\hat{\sigma}^2_e(z)} \qquad \text{....(27)}$$

अनुमानित घटक, प्रसरण के विश्लेषण (हन्डरसन विधि III,सरले इत्यादि 1992) उपरोक्त मॉडल को अनुप्रयुक्त कर प्राप्त किये जाते हैं।

सही पारिवारिक माध्य वंशागतित्व इस प्रकार है।

$$\hat{h}^2_{f(Z)} = \frac{\hat{\sigma}^2_f(z)}{\hat{\sigma}^2_f(z)+\hat{\sigma}^2_e(z)/n_{block}\times n_{plot}} \qquad \text{...(28)}$$

परिणाम और विचार–विमर्श

स्टेएबिलिटी की वंशागतित्व के अनुमान की दो विधियों की तुलना करने के लिए विभिन्न प्रकार के आँकड़े जिनकी विभिन्न प्रकार की असंतुलितता की मात्रायें है वह वंशागतित्व की विभिन्न प्राचल (parameter) के लिए कम्प्यूटर से अनुकरणित किये गये हैं।

Z_{ijk} पर आँकड़े रेखीय मॉडल के अनुसार से जनित हैं।

$Z_{ijk} = \mu + S_i + e_{ijk}$ एक सामान्यतः विस्तृत अनुपालनीय विचर Z के लिए जिसका पूर्ण (total) प्रसरण एक (1.0) है जोकि यादृच्छिक सम्पूर्ण ब्लॉकों में ऑफ–सिब की श्रेणी (series) में है।

परिवार मान (S_i) सामान्य प्रसरण की तरह अनुकरणीत है जिसका माध्य शून्य तथा प्रसरण 0.0125, 0.025, 0.0375 और 0.0625 है। त्रुटियां मानिकी वातावरणीय मान (e_{ijk}) एकल गोसियन विचर के रूप में अनुकरणीत है।

जिसका माध्य शून्य है तथा प्रसरण $(1-\sigma_f^2)$है। रूडन के पांच बिन्दुओं या प्रारम्भिक लैवलस वास्तविक आंकड़ों को द्विपद आंकड़ों में बदलने के लिए प्रयोग किया जाता है। वह प्रारम्भिक जिनका प्रयोग किया गया है। $p=0.05$, 0.010, 0.15, 0.20, 0.25 जोकि द्विपद विशेषक को अनुपालन करने की प्रायिकतायें हैं। आँकड़ों की **स्टेएबिलिटी** ($H_s^2=$ 0.05, 0.010, 0.15, 0.20, 0.25) की वंशागतित्व के विभिन्न प्राचल का प्रयोग कर जनित किया जाता है। प्राचलिता मान के लिए बीस जनकों के लिए नमूने जनित किये जाते हैं। जिनकी पुत्रियां पांच से चोबिस के बीच में होती हैं। इस प्रकार जनित अनुकरणित आंकड़ें **स्टेएबिलिटी** की वंशागतित्व के अनुमानों की विभिन्न प्रक्रियाओं के लिए अधीन है तथा इस प्रकार प्राप्त किये गये परिणाम तालिका–1 में दर्शित हैं तालिका–1 से यह देखा जाता है कि संकीर्ण संवेदीय बीटा–द्विपद संपादित वंशागतित्व ,$h^2_{rea(b)}$) किसी भी दूसरे अनुमानों से अच्छा परिणाम देती है। डैमस्टर लरनर अनुमान लगभग सामान्यतः प्रभावशाली है लेकिन परिवार माध्य वंशागतित्व निहायति अभिन्न है जानने के लिए यह एक रोचक बिन्दू है कि असंतुलिता के कारण मानक त्रुटि अत्यधिक बढ़ जाती है।

तालिका –1: असामान परिवार साइज के मामलों में दिये हुये h_s^2 (**स्टेएबिलिटी** की वंशागतित्व) कें विभिन्न मानों के लिए पशुओं के झुण्ड जीवन का व्यक्तिगत संकीर्ण संवेदीय वंशागतित्व(h^2) और परिवार माध्य वंशागतित्व(h_f^2)के औसत अनुमान।

vuqeku	h_S^2=0.05	h_S^2=0.10	h_S^2=0.15	h_S^2=0.20	h_S^2=0.25
h_Z^2	0.0502	0.1001	0.1503	0.2001	0.2450
	(0.0354)	(0.0525)	(0.0702)	(0.0848)	(0.0598)
$h^2_{rea(b)}$	0.0465	0.0987	0.1521	0.2092	0.2675
	(0.0671)	(0.0879)	(0.0879)	(0.1344)	(0.1600)
h^2_{DL}	0.0460	0.0977	0.1493	0.2045	0.2598
	(0.0660)	(0.0870)	(0.0961)	(0.1323)	(0.1570)
$h^2_{f(Z)}$	0.4105	0.6032	0.7031	0.7649	0.8086
	(0.2292)	(0.1560)	(0.1145)	(0.0905)	(0.0786)
$h^2_{f(beta)}$	0.1546	0.3310	0.4465	0.5306	0.5539
	(0.3098)	(0.2490)	(0.2085)	(0.1786)	(0.1554)
$h^2_{f(\Delta P)/beta)}$	0.1540	0.3295	0.4449	0.5206	0.5907
	(0.3084)	(0.2478)	(0.2056)	(0.2128)	(0.1547)

vkSlr ekud foPkyu czfdV esa gSa

vlarqfyrrk dh ekrzk = 35.0001.

उत्पादन के लिए समायोजन

प्राचलित(पैरामैट्रिक) मान h_Y^2 का प्रयोग करके **स्टेएबिलिटी** के लिए आँकड़ें (उत्पादन की वंशागतित्व) =0.25, m_Y (उत्पादन पर पशुओं के झुण्ड जीवन के मानविकृत समाश्रयण गुणांक)=0.4, $r_{Y,HL}$ (पशुओं के झुण्ड जीवन और उत्पादन दृश्य प्रारूप सहसंबंध)=0.25, इन प्राचल मानों के लिए तर्क जिसके साथ–साथ **स्टेएबिलिटी** की विभिन्न वंशागतित्व (h_S^2=0.05,0.10,0.15,0.20,0.25) भी डैकर्स ;1993द्ध से लिया गया है। इस प्रकार प्राप्त समायोजित आँकड़ें आगे चल कर द्विपद स्कैल में बदल दिया जाता है। समायोजित **स्टेएबिलिटी** आँकड़ों के लिए परिणाम तालिका–2 में दिखाया गया है। डैमस्टर लरनर तथा संकीर्ण संवेदीय बीटा–द्विपद वंशागतित्व अनुमानों के मामलों में बेहतर परिणाम देखे गये हैं। बाकी दूसरे प्रक्रिया के लिए परिणाम अत्याधिक अभिनत है। यह देखना बहुत रोचक है कि समायोजिता के कारण ना सिर्फ अनुमान जनसंख्या प्राचल के बहुत नजदीक थे। बल्कि त्रुटियॉ अत्याधिक कम हो गयी और इस प्रकार अनुमानो की यथार्थता बढ़ा दी।

तालिका –2: असामान परिवार साइज के मामलों में दिये हुये h_s^2 (**स्टेएबिलिटी** की वंशागतित्व)के विभिन्न मानों के लिए उत्पादन के लिए समायोजित पशुओं के झुण्ड जीवन की व्यक्तिगत संकीर्ण संवेदीय वंशागतित्व $\left(h^2\right)$ और परिवार माध्य वंशागतित्व $\left(h_f^2\right)$ के औसत अनुमान।

Estimate	h_S^2=0.05	h_S^2=0.10	h_S^2=0.15	h_S^2=0.20	h_S^2=0.25
h_Z^2	0.0503	0.0977	0.1457	0.1937	0.2420
	(0.0334)	(0.0484)	(0.0615)	(0.0751)	(0.0884)
$h_{rea(b)}^2$	0.0524	0.1027	0.1536	0.2079	0.2628
	(0.0679)	(0.0872)	(0.1050)	(0.1278)	(0.1503)
h_{DL}^2	0.0534	0.1027	0.1529	0.2058	0.2586
	(0.0681)	(0.0868)	(0.1075)	(0.1320)	(0.1563)
$h_{f(Z)}^2$	0.5113	0.7056	0.7912	0.8393	0.8702
	(0.2951)	(0.1451)	(0.0929)	(0.0676)	(0.0528)
$h_{f(beta)}^2$	0.1866	0.3488	0.4550	0.5338	0.5926
	(0.3015)	(0.2449)	(0.2041)	(0.1745)	(0.1541)
$h_{f(\Delta P)/beta)}^2$	0.1857	0.3457	0.4529	0.5313	0.5899
	(0.3002)	(0.2482)	(0.2031)	(0.1733)	(0.1544)

अkSlr ekud foPkyu czfdV esa gSa

vlarqfyrrk dh ekrzk = 35.0001.

रूट माध्य वर्ग त्रुटि

विभिन्न विधियों की अनुभाविक तुलना के लिए औसत रूट माध्य वर्ग त्रुटि बहुत उपयोगी पायी गया है तथा विभिन्न वंशागतित्व और विभिन्न प्रारिम्भक प्रायिकताओं पर परिकल्पना की गयी है। बीस जनको जिसमें पाँच से चौबीस पुत्रियों और जो ब्लाक साईज़ पॉच में से उनके लिए रूडन प्रायिकताओं औसत की गई आपेक्षिक रूट माध्य वर्ग त्रुटि तालिका–3 में दिखायी गयी है और **स्टेएबिलिटी** की वंशागतित्वों के विभिन्न मानों पर समान औसत तालिका–4 में दिखायी गयी है तालिका–3 तथा तालिका–4 से यह साफतौर पर देखा गया है कि परिवार माध्य वंशागतित्व अनुमानों के लिए आपेक्षित रूट माध्य वर्ग त्रुटि बाकी किसी भी वंशागतित्व अनुमान से संख्यात्मकता ज्यादा सार्थकता पूर्ण है। तालिका–4 से यह देखा गया है कि सही आँकड़ों बिन्दुओं पर परिवार माध्य वंशागतित्वों की आपेक्षिक रूट माध्य वर्ग त्रुटि उच्चतम मान है।

तालिका–3 असमान पुत्री के मामलों में पशुओ के झुण्ड जीवन की वंशागतित्व के चयनित अनुमानों की आपेक्षिक रूट माध्य वर्ग त्रुटि (RMSE%)

vuqeku	$h^2_{rea(b)}$	h^2_{DL}	$h^2_{f(beta)}$	$h^2_{f(\Delta P)/beta)}$	h^2_Z	$h^2_{f(Z)}$
h^2_S=0.05	134.8040	132.2291	670.1510	666.942	70.9687	854.3492
	(131.040)	(136.600)	(681.601)	(678.257)	(66.697)	(1055.292)
h^2_S=0.10	87.9778	85.5209	351.2075	349.3008	52.5299	526.8245
	(86.012)	(86.945)	(359.704)	(310.002)	(47.592)	(622.237)
h^2_S=0.15	72.7759	70.7717	248.2335	206.4008	46.8208	376.5623
	(70.074)	(71.692)	(250.656)	(249.152)	(41.091)	(431.909)
h^2_S=0.20	67.3535	66.1898	192.4388	199.7637	42.3125	286.0470
	(64.049)	(64.866)	(191.858)	(190.557)	(37.681)	(321.4421)
h^2_S=0.25	64.3965	62.9858	154.3059	152.6061	39.9093	224.4113
	(60.3965)	(62.637)	(150.892)	(150.893)	(35.516)	(248.963)

Lkek;ksftrk के मामलों में समरूप रूट माध्य वर्ग त्रुटि ब्रैकिट में है।

तालिका 4:– असमान पुत्रियों के मामले में पशुओ के झुण्ड जीवन में अनुमानों की आपेक्षिक रूट माध्य वर्ग त्रुटि(RMSE%)

Estimate	$h^2_{rea(b)}$	h^2_{DL}	$h^2_{f(beta)}$	$h^2_{f(\Delta P)/beta)}$
$\bar{p}$ =0.05	118.3930	117.7272	349.6306	348.0157

	(117.765)	(125.393)	(353.457)	(350.779)
$\bar{p}$ =0.10	91.169	89.4294	334.3121	332.5418
	(87.809)	(91.196)	(346.205)	(342.534)
$\bar{p}$ =0.15	72.4481	77.3592	329.7966	328.0628
	(76.218)	(75.920)	(335.362)	(336.125)
$\bar{p}$ =0.20	71.3920	68.1943	301.6859	299.0613
	(69.549)	(67.924)	(304.375)	(302.457)
$\bar{p}$ =0.25	67.6236	64.9872	300.3517	298.4655
	(65.008)	(62.305)	(299.289)	(296.969)

समायोजिता के मामलों में समरूप रूट माध्य वर्ग त्रुटि ब्रैकिट में है।

समायोजन के मामले में रूट माध्य वर्ग त्रुटि में सार्थक बदलाव नहीं देखा गया। **स्टेएबिलिटी** के अनुवांशिक मानों तथा रूडन बिन्दुओं में बढ़ते हुये चलन से आपेक्षिक बिन्दु के लिए वंशागतित्व के अनुमानों की आपेक्षिक रूट माध्य अनुमानों से अनुसरीत है। सही मानों तथा बीटा–द्विपद प्रक्रिया के परिणामों के अनुमानों तो कुछ एकरूपता दिखाते हैं। जबकि दूसरी प्रक्रिया में उत्पादन लक्षण पर आधारित आंकड़ों के समायोजन के लिए कोई विशेष ट्रेन्ड नहीं देखा गया। इन परिणामों से अन्ततः यह निष्कर्ष निकलता है अगर किसी को **स्टेएबिलिटी** के आँकड़े निर्धारित करने का प्रस्ताव हो तो इस निर्धारित आँकड़ों पर आधारित वंशागतित्व के अनुमान बहुत ही दक्षतापूर्ण तथा स्पष्ट अनुमान देगा और यदि किसी के पास **स्टेएबिलिटी** के सिर्फ द्विगुण आँकड़े हो तब बीटा–द्विपद अनुमान बाकी अनुमानों की विधियों की तुलना में बहुत ही सटीक और स्पष्ट हैं। जबकि असंतुलिता कई बार अनुमानों में बड़ी मानक त्रुटियों की वजह बन जाती है।

संदर्भ

1. डेक्कर्स जैक, सी.एम., 1993. थियोरेटिक्ल बेसिस फॉर जेनेटिक पैरामीटर ऑफ हर्ड लाईफ एण्ड इफेक्टस ऑन रिसपोंस टू सलेक्शन. जे. डेयरी सांईस, 76: 1433–1443.
2. डेम्पसटर, ई.आर. एण्ड लर्नर, आई.एम., 1950. हैरिटेबिल्टि ऑफ थ्रशोल्ड करेक्टर्स. जेनेट., 35: 212–236.
3. फालकोनर, डी.एस., 1981. इन्ट्रोडक्शन टू क्वानटीटेटिव जेनेटिक्स, सेकेन्ड एडीशन, लांगमैन, लंदन.
4. गियानोला, डी., 1979 हैरीटेबिल्टि आफ पोलीकोटोमस करेक्टर्स. जेनेट., 93: 1051–1055.
5. जॉनसन, एन.एल. एण्ड कोटज़ एस., 1970. कनटिनयूअस यूनीवेरीयेट डिस्ट्रीब्यूशन, 2. जॉन विले एण्ड सन्स, न्यूयार्क.
6. मैगनुसेन. एस. एण्ड क्रिमर, ऐ. 1995. दी बीटा–बायनोमियल मॉडल फार एस्टीमेटिंग आफ बाइनरी ट्रेटस. थीयोरेटि. एपीली. जैनेट., 91: 544–552.
7. सरले, एस.आर. कासीला. जी. एण्ड मैक्कुलाच, सी.ई., 1992. वेरियेन्स कम्पोनेंटस. जॉन विले एण्ड सन्स, न्यूयार्क.

मैटिंग डिज़ाइन तथा एनवायर्नमेंटल डिज़ाइन

डॉ. सिनी वर्गीस
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

1- çLrkouk

ikni ,oa i'kq çtuu dk;ZØeksa dk ,d çeq[k mís';] ikS/kksa ,oa i'kqvksa dh vkuqoaf'kd {kerk esa lq/kkj djuk gSA çtuu& ç;ksxksa esa nks çdkj ds fMtkbuksa uker% esfVax fMtkbuksa ,oa i;kZoj.kh; fMtkbuksa dk lekos'k gksrk gSA esfVax fMtkbu larfr;ksa dks mRiUu djus dh ,d ç.kkyhxr çfd;k gSk bu larfr;ksa dh ç.kkyhxr <ax ls i;kZoj.kh; ifjfLFkfr;ksa esa j[kj[kko esa o`f) ds fy, i;kZoj.kh; fMtkbu dk mi;ksx gksrk gSA Mkb,yhy Ø,l] vkaf'kd Mkb ,yhy Økl] Vªkb ,yhy ,oa vkaf'kd Vªkb,yhy Ø,l] rFkk Mcy Ø,l o vkaf'kd VsVªk ,yhy Ø,l] lkekU;r% mi;ksx esa yk, tkus okys esfVax fMtkbu gSaA

buesa ls Mkb ,yhy ,oa vkaf'kd Mkb ,yhy Ø,l ¼ihMhlh½ ;kstuk,a] lokZf/kd yksdfç; gSaA Mkb ,yhy Ø,l] dbZ thuç:iksa ds chp lHkh laHko esfVax dk ,d lSV gS tks ,dy] Dyksu] lekaxh oa'kØe ¼ykbus½ vkfn gks ldrs gSaA ikni ,oa i'kq çtuu] nksuksa esa ,d ek=kRed xq.k ds ldy çlj.k ds vkuqoaf'kd ?kVdksa dks Kku djus ds fy, bu Ø,lksa dk ckjEckj mi;ksx fd;k tkrk gSA Ø,lksa esa lfEefyr var:çtkr oa'kØeksa dh fof'k"V ;kstu ;ksX;rkvksa ds fu/kkZj.k esa Hkh budk mi;ksx gksrk gSA O;qRØe Øklksa ,oa tud var:çtkrksa dks NksM+dj] oa'kØeksa ds ,d lSV esa अ;अ.1द्ध2 laHko Ø,l gksrs gSaA v esa c<+ksrjh ds lkFk Ø,lksa dh la[;k esa rsth ls c<+ksrjh gksrh gSA mnkgj.kkFkZ] gksus v ¾ 6 gksus ij ;s 15] v ¾20 gksus ij ;s 190 rFkk v ¾ 50 gksus ij budh la[;k 1225 gks tkrh gSA ;fn ijh{k.k djus dh lqfo/kk,a lhfer gksa rks vis{kk—r de la[;k esa var% çtkr oa'kØeksa gsrq ,d Mkb,yhy Ø,l gh laHko gSA

ihMhlh ds ek/;e ls] ,d ikni iztud u dsoy ,d cM+h la[;k esa tud oa'kdzeksa dk thlh, Kkr dj ldrk gS cfYd tudksa dh ,d O;kid lhek rd dzkWlksa ds chp oj.k Hkh dj ldrk gSA izR;sd oa'kdze dk thlh, vis{kkd`r de ifj'kq)rk ls Kkr gksrk gS fdarq mlds lkFk vf/kd l?ku oj.k dk vuqiz;ksx ifj.kkeLo:i vf/kd vkuqoaf'kd ykHk izkIr gks ldrk gSA ;g lwfpr djuk egRoiw.kZ gS fd fdlh ifjfLFkfr esa] ihMhlh gsrq dbZ laHko fMtkbuksa esa ls dkSu lk bu n`f"V ls lokZf/kd l{ke gS fd og oa'kdzeksa ds ,d ;qXe ds thlh, izHkkoksa ds chp fHkUurk dk U;wure vkSlr izlj.k nsrk gSA

1- ldqZysaV lajpuk ij vk/kkfjr ihMhlh ;kstuk,a

fgadsyeSu ,oa LVuZ ¼1960½ us dqN ldqZysaV uewuksa dk o.kZu fd;k gS ftuesa oa'kdze 1 dks ges'kk oa'kdze 2 ,oa mu oa'kdzeksa ds lkFk dzkWl fd;k tkrk gS ftudh la[;k 2 ls N rd ,d vad xf.krh; vuqdze cukrh gSA mnkgj.kksa dk izLrqrhdj.k] N = 8, s = 3 ,oa N =

14, s = 5 ds fy, fd;k x;k FkkA dsEiFkzksu ,oa djukm ¼1961½ us ihMhlh ds ldqZysaV uewus cukus dh fof/k;ka nh Fkha] pkgs N le Fkh ;k fo"ke tSls fd N+ s fo"ke iw.kkZad gSA

ekuk ,d iztud] dqy ns/2 dzklksa ds lkFk dk;Z dj ldrk gS tgka n variztkr oa'kdzeksa dh la[;k gS vkSj s ,d iw.kkZad **2 ≥** gSA n variztkr ;kn`fPNd :i ls] 1 ls ysdj n rd gSa vkSj dzkWlksa ds uewus fuEufyf[kr gSa%

oa'kdze 1 × oa'kdze k+1, k+2, …, k+s
oa'kdze 2 × oa'kdze k+2, k+3, …, k+1+sVII-2
oa'kdze i × oa'kdze k+i, k+i+1, …, k+i-1+s
oa'kdze n × oa'kdze k+n, k+n+1, …, k+n-1+s ⋮⋮

tgka k=(n+1-s)/2 rFkk s og la[;k gS ftruh ckj ,d oa'kdze dzkWlksa esa lfEefyr gksrh gSA n ls Åij dh la[;kvksa dks eksM~;wyks n rd de dj fn;k tkuk pkfg, vkSj k ds fy, ,d iw.kkZad rd de dj nsrs gSaA Li"Vr;k] n ,oa s nksuksa fo"ke la[;k ugha gks ldrhA

;fn s le gS fdarq n ds le ekuksa gsrq s fo"k; gS rks vk;Z ¼1983½ us lHkh n ds fy, ihMhlh izkIr djus ds fy, la'kksf/kr ldqZysaV ;kstukvksa dk lq>ko fn;k gSA

3- vkaf'kd :i ls larqfyr viw.kZ CykWd fMtkbu dh ,slksfl,'ku Ldhe ij vk/kkfjr ihMhlh ;kstuk,a

ihMhlh cukus esa ,d vkaf'kd :i ls larqfyr viw.kZ CykWd fMtkbu dh ,lksfl,'ku Ldhe ¼ihchvkbZch½ fMtkbu dk lQyrkiwoZd iz;ksx fd;k tk ldrk gSA fxycVZ ¼1958½ us ihMhlh ,oa lkbt nks ds CykWDl ds chp vuq:irk dh vksj bafxr fd;k gSA bldh O;k[;k djus ds fy, ,d 2 ,lksfl,V Dykl ihchvkbZch ,lksfl,'ku Ldhe dk mi;ksx fd;k tk ldrk gSA V mipkjksa gsrq ,d viw.kZ CykWd fMtkbu dks m& ,lksfl,V Dyklst ds lkFk vkaf'kd :is.k larqfyr ekuk tkrk gS] ;fn izk;ksfxd lkexzh dks ,sls b CykWDl esa foHkkftr fd;k tk ldrk gks ftues als izR;sd dk lkbt k (< v) gks tSls fd &

¼i½ izR;sd mipkj r CykWDl esa gksrk gS]
¼ii½ mipkjksa ds chp ,d ,ClVªSDV laca/k gks tks fuEufyf[kr ckrksa dks iwjk djrk gks
¼d½ nks mipkj] 1^{st}, 2^{nd}, … m^{th} ,lksfl,V gS] ,lksfl,'ku dk laca/k lefrrh; gS vFkkZr ;fn mipkj α, β dk i^{th} ,lksfl,V gS rks β, α dk i^{th} ,lksfl,V Hkh gS ¼i = 1,2,…,m½]
¼[k½ izR;sd mipkj esa fuf'pr :i ls n_i i^{th} ,lksfl,V~l gksrs gSa vkSj
¼x½ fn, x, dksbZ Hkh nks mipkj tks ijLij i^{th} ,lksfl,V~l vkSj nwljs ds j^{th} ,lksfl,V~l vkSj nwljs ds k^{th} ,lksfl,V~l ds fy, mipkjksa dh dkWeu la[;k pijk (i,j,k = 1,2,…,m).
¼iii½ nks mipkj tks ijLij i^{th} ,lksfl,V~l gSa] lkFk&lkFk fuf'pr :i ls λi CykWDl esa gksrs gSaA

v, b, r, k, n_i, λ_i la[;k,a izFke izdkj ds izkpy dgykrs gSa tcfd la[;k,a pijk (i,j,k = 1,2,3,…,v) nwljs izdkj ds izkpy dgykrs gSaA izR;sd mipkj dks Mkb,yhy dzkWl esa lfEefyr ,d oa'kdze le>k tk ldrk gSA vc izR;sd oa'kdze ds lkFk blds izFke ,lksfl,V~l ;k] nwljs ,lksfl,V~l ;k] rhljs ,lksfl,V~l ds lkFk Bhd ,d ckj lHkh laHko dzkWlst cuk,aA dzkWlst dk izR;sd lSV] ,d ihMhlh ;kstuk cukrk gSA blfy, ,d m- ,lksfl,V Dykl ,lksfl,'ku Ldhe mihMhlh ;kstuk,a izkIr gksrh gSa vkSj os lHkh lqLi"V gksrh gSaA rFkkfi] ,d dusDVsM Iyku dk mi;ksx djrs le; lko/kkuh cjruh pkfg, rkfd thlh, izHkkoksa esa izkFkfed fojks/kkHkklksa ls lacaf/kr izlj.kksa dks Kkr fd;k tk ldsA

v=10 mipkjksa dks ,d 2& ,lksfl,V Dykl f=dks.kh; ,lksfl,'ku Ldhe esa bl izdkj ls O;ofLFkr fd;k tkrk gSA

*	1	2	3	4
1	*	5	6	7
2	5	*	8	9
3	6	8	*	10
4	7	9	10	*

;gka nks mipkj izFke ,lkssfl,V gksrs gSa ;fn og ,js ds leku iafDr ;k dkWye esa mRiUu gksrs gSa] vU;Fkk ;g f}rh; ,lkssfl,V gksrs gSaA vc mipkjksa dks oa'kdzeksa ds :i esa fy;k x;k vkSj izR;sd oa'kdzeksa ds chp lHkh laHkkfor dzklksa dk l`tu djrs gq, vkSj blds f}rh; ,lksf'k,V dks Li"Vr;k ,d ckj ysrs gq, ,d ihMhlh ;kstuk esa 15 dzklksa dks 'kkfey ¼leLr 45 dzklksa esa ls½ djrs gq, izkIr dzklksa dks uhps fn;k x;k gS %

1×8 1×9 1×10
2×6 2×7 2×10
3×5 3×7 3×9
4×5 4×6 4×8
5×10 6×9 7×8

ckn esa ?kks"k vkSj nsfopk ¼1997½ us ihchvkbZch ls ihMhlh ds lkFk&lkFk f}rh; lacaf/kr oxksZa dks O;qRiUu fd;k gSA f}rh; lacaf/kr oxZ ,lksfl,'ku Ldhe dk bLrseky djrs gq, ,d ihMhlh esa 'kkfey dqy dzklksa dh la[;k dkQh vf/kd gks ldrh gSA blds QyLo:i bu lHkh dks izHkko'kkyh <ax ls laHkkyuk eqf'dy gksrk gSA QkbQ vkSj fxycVZ ¼1963½ us rhu ,lksfl,V Dykl ds lkFk ihchvkbZch fMtkbu ls O;qRiUu ihMhlh dks lekfo"V fd;kA nkl vkSj f'kojke ¼1968½ us fdlh Hkh CykWd fMtkbu] h dh fdlh Hkh oSY;w rFkk ,lksfl,V Dykl dh fdlh Hkh la[;k ds lkFk ihchvkbZch fMtkbu dk bLrseky djrs gq, vkaf'kd Mkb ,yhy dzkWl gsrq ;kstuk cukbZA fgadyeku rFkk dsaisFkkZu us ihMhlh Iyku dks rS;kj djus esa ;ksxnku fn;k D;ksafd lewg ds nks ,lksfl,V Dykl foHkkT; ihMhlh FksA ukjk;.k vkfn ¼1974½ us foLrkfjr f=Hkqtkdkj ,lksfl,'ku Ldhe ds vk/kkj ij ihMhlh ds fo'ys"k.k esa lg;ksx fn;k gSA vusd ,lksfl,'ku Ldhe ds vk/kkj ij ihMhlh ds l`tu vkSj fo'ys"k.k ds {ks= esa vk;kZ vkSj ukjk;.k ¼1977½] ukjk;.k rFkk vk;kZ ¼1981½] vxzoky ¼1985½] dkSf'kd vkSj iqjh ¼1989½ rFkk dkSf'kd ¼1999½ }kjk vkSj T;knk

;ksxnku fn;k x;k gSA dqN ihMhlh lyku dks rhu ,lksfl,V Dykl ds lkFk ihchvkbZch fMtkbu ls izkIr fd;k x;k vkSj lkfgR; esa miyC/k lwphdj.k dks oxhZt vkfn ¼2005½ esa ns[kk tk ldrk gS ¼2005½A

mnkgj.k% V=12 mipkjksa dks fuEufyf[kr :i n'kkZ, x, rjhds ls jsDVsaxqyj ,lksfl,'ku Ldhe ¼3 ,lksfl,V Dykl ihchvkbZ fMtkbu½ esa O;ofLFkr fd;k tk ldrk gS%

1	2	3
4	5	6
7	8	9
10	11	12

fuEufyf[kr rkfydk esa igys 6 mipkjksa ds rhu vyx&vyx ,lksfl,V dks n'kkZ;k x;k gS%

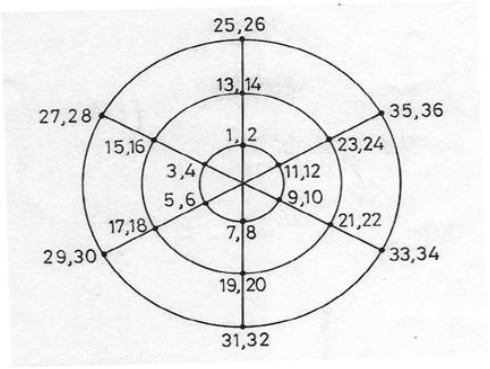
mipkj	1st ,lksfl,V~l	2nd ,lksfl,V~l	3rd ,lksfl,V~l
1	2, 3	4, 7, 10	5, 6, 8, 9, 11, 12
2	1, 3	5, 8, 11	4, 6, 7, 9, 10, 12
3	1, 2	6, 9, 12	4, 5, 7, 8, 10, 11
4	5, 6	1, 7, 10	2, 3, 8, 9, 11, 12
5	4, 6	2, 8, 11	1, 3, 7, 9, 10, 12
6	4, 5	3, 9, 12	1, 2, 7, 8, 10, 11

bl mnkgj.k dk mi;ksx djrs gq, geus 12 oa'kdzeksa esa fuEu rhu ihMhlh lyku izkIr fd, gSaA

lyku D_1	**lyku D_2**	**lyku D_3**
1×2, 1×3, 2×3, 4×5, 4×6, 5×6, 7×8, 7×9, 8×9, 10×11, 10×12, 11×12	1×4, 1×7, 1×10, 2×5, 2×8, 2×11, 3×6, 3×9, 3×12, 4×7, 4×10, 5×8, 5×11, 6×9, 6×12, 7×10, 8×11, 9×12	1×5, 1×6, 1×8, 1×9, 1×11, 1×12, 2×4, 2×6, 2×7, 2×9, 2×10, 2×12, 3×4, 3×5, 3×7, 3×8, 3×10, 3×11, 4×8, 4×9, 4×11, 4×12, 5×7, 5×9, 5×10, 5×12, 6×7, 6×8, 6×10, 6×11,

		7×11, 7×12, 8×10, 8×12,9×10, 9×11

mnkgj.k% v = 36 $(=2\times2\times3^2)$ ladsfUnzr o`Rr rFkk 3 Mk;kehVj ds 18 loZfu”B fcanqvksa ij 36 mipkjksa dks bl izdkj O;ofLFkr fd;k x;k ftlesa izR;sd fcanq esa nks mipkj ‘kkfey Fks tSlk fd uhps n’kkZ;k x;k gS%



mipkj 1 ds ,lksfl,V fuEuor gS

izFke ,lksfl,V~l	f}rh; ,lksfl,V~l	r`rh; ,lksfl,V~l
2, 7, 8	3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 19, 20, 25, 26, 31, 32	15, 16, 17, 18, 21, 22, 23, 24, 27, 28, 29, 30, 33, 34, 35, 36

ihMhlh lyku dks leku:ih rjhds ls izkIr fd;k tk ldrk gSA

;fn] fn;k x;k V ¼oa’kdze la[;k½ mPp ,lksfl,V Dykl ds lkFk ,d fMtkbu gS rks ladjksa dh la[;k de gksxh vkSj ,lksfl,V ds fodYi esa T;knk yphykiu gksxkA

mnkgj.k% nh xbZ oa’kkofy;ksa dh la[;k ds fy,] mnkgj.k% V=12 dks izR;sd vkdkj 4 ds 3 xzwiksa esa O;ofLFkr fd;k x;k] ;fn ge nks Dykl xzqi&foHkkT; ,lksfl,’ku Ldhe ds fy, vkxs c<+rs gS] gesa nks izdkj ds ladj feysaxs% 18 ladjksa esa igys ,lksfl,V ifj.kke ds lkFk izFke vkSj nwljs ,lksfl,V ds lkFk nwljk ftlds QyLo:i 48 dzkWl Fks fdarq leku:ih V ds lkFk] ;fn bl 4 ,lksfl,V Dykl lgh dks.hk; ,lksfl,’ku Ldhe ds lkFk vkxs c<+rs gSa rks gesa 4 fofo/k izdkj ds ,lksfl,V ds lkFk 4 fofHkUu izdkj ds dzkWl fey ldrs gSa blds QyLo:i dze’k% 6 dzkWl] 12 dzkWl] 24 dzkWl rFkk 24 dzkWl rd gks ldrs gSaA

dq’kyre esfVax fMtkbu

izR;sd oa’kkoyh ds ladj.k ds lkFk blds m ,lksfl,V ¼m=1,2,3,… ½ }kjk esfVax fMtkbu izkIr djus ds fy, m- ,lksfl,V Dykl ihchvkbZch fMtkbu dh ,lksfl,’ku Ldhe dk mi;skx fd;k tk ldrk gSA

fuEufyf[kr vyx&vyx lyku izkIr fd, x,%

lyku D1: V oa’kkoyh Fkk buds igys ,lksfl,V ds lkFk ladj.k ls n2v/2 dzkWl izkIr gksrs gSa] lyku D2: V oa’kkoyh dk buds f}rh; ,lksfl,V ds lkFk ladj.k ls nmv/2 dzkWl izkIr gq,A bu lykuksa esa ls izfr dzkWl dh vf/kdre lwpuk dk p;u fd;k x;kA

vkdfyr varj (gˆ i -gˆi’) dh tkap ds fy, m izlj.k gksxk D;ksafd oa’kkoyh i rFkk i ’ izFke] f}rh;] ;k ---------- m^{th} ,lksfl,V gS] tks izR;sd Dm, m = 1, 2, 3,.. gS

m^{th} lyku ds fy, gca izHkkoksa ds chp varj dk vkSlr izlj.k gS

Page VII-5 formuls

m^{th} lyku ls izfr dzkWl izkIr lwpuk gS

Page VII-5 formuls

laiw.kZ Mkb ,yhy dzkWl lyku dh rqyuk esa izfr dzkWl lwpuk ds lanHkZ esa n{krk ?kVd] nksuksa lyku ds fy, =qfV izlj.k dks leku ekuk x;k] ;g fuEufyf[kr :i esa ik;k x;kA

Page VII-5 formuls

4- **daI;wVj }kjk l`ftr b”Vre ihMhlh lyku**

vuqdwyrk ekunaM] uker% Vªsl dzkbVsfj;k ij vk/kkfjr vuqdwy lyku dk irk yxkrs gq, leLr fofo/k dks.kksa [ekFkqj vkSj ukjk;.k] 1976] Mk,yhy dzkWl uewuksa dh leL;k dk irk yxk;k tk ldrk gSA N=n(n-1)/2 dzkWl esa ls ,d b = ns/2 dzkWl dk mi;ksx bl izdkj fd;k tk ldrk gS tks ;FkklaHkkfor thlh, izHkko dh rqyuk esa izlj.k dk vkSlr gSA ;g ihMhlh ds fy, vuqdwy izHkko dks c<+krk gSA

mnkgj.k% izR;sd ds lkFk n=4 oa'kdzeksa ij fopkj djrs gq, vU; oa'kdzeksa ds lkFk s=4 ckj dzkl fd;k x;kA gekjs ikl N = 36 rFkk b = 18 gSA bl izdkj $^{36}C_{18}$ lyku dks daI;wVj ij l`ftr fd;k tk ldrk gSA vkdyu (gi –gjˆ½ ds U;wure vkSlr izlj.k dks b"Vre lyku ds :i esa fy;k tk ldrk gSA

5- esfVax & i;kZoj.kh; fMtkbu

i`Fkd esfVax fMtkbu ds cxSj vkxs c<+us ds fy, la;qDr esfVax i;kZoj.kh; fMtkbu dks pquk x;k vkSj mfpr i;kZoj.kh; fMtkbu ¼ihchvkbZch fMtkbu½ ds CykWd ds rgr lHkh laHkkfor dzkWl }kjk uewus izkIr fd, x,A

vxzoky vkSj nkl ¼1990½] flag vkSj fgadyeku ¼1995½ rFkk 'kekZ ¼1998½ ds ihMhlh ijh{k.kksa ds fy, v/kwjs CykWd fMtkbu ds bLrseky dk izn'kZu fd;kA

mnkgj.k% ekuk fd 9 mipkj gSaA bu mipkjksa dks o`Rrkdkj ?ksjs esa 9 arcs esa foHkDr djds bl izdkj O;ofLFkr fd;k gS ftles a,d mipkj esa izR;sd dks 'kkfey fd;k x;kA

page VII 6 photo

v = 9 ds lkFk fuEufyf[kr o`Rrkdkj fMtkbu dks b=9 CykWd esa O;ofLFkr fd;k x;k] rhu fu;fer arcs ds la;qDr ?kVd }kjk izR;sd vkdkj 3 dks izkIr fd;k tk ldrk gS%

1	2	3
2	3	4
3	4	5
4	5	6
5	6	7
6	7	8
7	8	9
8	9	1
9	1	2

izR;sd CykWd esa lHkh laHkkfor fof'k"V dzkWl ds xBu }kjk ,d la;qDr esfVax i;kZoj.kh; fMtkbu dks izkIr fd;k tk ldrk gS tks fuEufyf[kr gS%

1×2	1×3	2×3
2×3	2×4	3×4

3×4	3×5	4×5
4×5	4×6	5×6
5×6	5×7	6×7
6×7	6×8	7×8
7×8	7×9	8×9
8×9	8×1	9×1
9×1	9×2	1×2

6- iSr`d oa'kdzeksa ds nks lewgksa dks 'kkfey djrs gq, laof/kZr vkaf'kd Mkb ,yhy dzkWl ;kstuk,a

dqN izk;ksfxd fLFkfr;ksa esa] iz;ksxdrkZ ds ikl iSr`d oa'kdzeksa ds nks lewg gks ldrs gSa] ,d lewg esa oSls oa'kdze gksrs gSa tks izkFkfed egRo ds gSa vkSj nwljk lewg f}rh;d :fp ds oa'kdzeksa ls fufgr gksrk gSA ,d iztud dk;Zdze esa dzklksa dks ,d mPp lekuqikr esa mRd`"V ;k lq&vuqdwfyr oa'kdzeksa dk izfrfuf/kRo nsuk ges'kk gh okaNuh; gksrk gSA bls vftZr djus ds fy,] izR;sd izkFkfed oa'kdze dk 'ks"k nwljs oa'kdzeksa ds lkFk ladj.k fd;k tkuk pkfg, tcfd f}rh; oa'kdze ds chp dsoy ,d ihMhlh gksuk pkfg,A ,slh fLFkfr;ksa ds fy, esfVax fMtkbuksa dks laof/kZr vkaf'kd Mkb ,yhy dzkWl ¼,ihMhlh½ dgk tkrk gS tks fd lhMhlh vkSj ihMhlh dk la;kstu gSA bldk mi;ksx ,slh fLFkfr;ksa esa fd;k tkrk gS tgka ,d lhMhlh O;ogk;Z ugha gS vkSj iz;ksx esa f}rh;d oa'kdzeksa dh rqyuk esa izkFkfed oa'kdzeksa ds ckjs esa vf/kd tkudkjh izkIr dh tkuh gSA

isMlZu ¼1980½ us oa'kdzeksa ds thlh, ,oa ,llh, izHkkoksa ds vkdyu ds fy, ,ihMhlh dh fMtkbu rS;kj dhA osadVslu ¼1985½ us ihMhlh iz;ksxksa ds yphys izdkj ij fopkj fd;k ftlesa oa'kdzeksa ds ,d lewg dk ladj.k bl vuqikr ds lkFk vU; nwjLFk fLFkr lewg ds oa'kdzeksa ds lkFk fd;k x;k fd izR;sd lewg ds Hkhrj ladj.k ,d izkjafHkd pj.k esa gksA tXxh ,oa vxzoky ¼1995½ us ,ihMhlh ;kstukvksa dk ,d iz.kkyhxr fo'ys"k.k fodflr fd;kA tXxh ,oa 'kqDyk ¼1996½ us ,ihMhlh ,oa lhMhlh ;kstukvksa ds chp ,d rqyuk dhA dqfj;kdkst ¼1998½ us laof/kZr Mkb ,yhy dzkWl ;kstuk cukbZ ftlesa larqfyr viw.kZ CykWd ¼chvkbZoh½ fMtkbu CykWd dk mi;ksx djrs gq, izkFkfed ,oa f}rh;d oa'kdzeksa ds thlh, dh rqyuk ds fy, larqfyr fd;k tk ldsA JhokLro ¼2010½ us vkaf'kd :i ls larqfyr viw.kZ CykWd fMtkbuksa dh fofHkUu laxr ;kstukvksa dk mi;ksx djrs gq, ,ihMhlh ;kstukvksa ds dqy oxksZa dks izkIr fd;kA

,ihMhlh ;kstuk ds fuekZ.k dh dk;Zfof/k

eku yhft, izkFkfed ,oa f}rh;d oa'kdzeksa dh la[;k dze'k% p ,oa q gS ftlesa p + q = N gSA ,ihMhlh ladj.k iz.kkyh esa] ,slk ekuk tkrk gS fd izR;sd izkFkfed oa'kdze dk vU; lHkh oa'kdzeksa ds lkFk dzkWl djk;k tkrk gS ftlls ¼N-1½ ladj.k izfr izkFkfed oa'kdze mRiUu gksrk gSA blds vfrfjDr] m ,lksfl,V oxZ PBIB fMtkbu dh laxr ;kstukvksa dk mi;ksx djrs gq, f}rh;d

oa’kdzeksa ds chp ,d vkaf’kd Mkb ,yhy dzkWl fd;k tkrk gS blds n_i i^{th} ,lksfl,V~l ds lkFk izR;sd f}rh;d oa’kdze ds chp ladj.k ls qn_i $(i = 1, 2, \ldots, m)$ ladj mRiUu gksrs gSaA blfy,] jsflizksdy ¼O;qRdzeksa½ dks NksM+dj] ,ihMhlh ;kstuk ¼N ,ihMhlh½ ds fy, dzkWlksa dh dqy la[;k $[p(N - 1) + q(p + n_i)]/2$ gSA

,ihMhlh esfVax fMtkbu

,ihMhlh fMtkbuksa ds fuekZ.k ds fy, fofHkUu laxr ;kstukvksa dk mi;ksx fd;k tk ldrk gS%
¼i½ f=dks.kh; laxr ;kstuk
¼ii½ lewg foHkkT; laxr ;kstuk
¼iii½ ySfVu LDok;j laxr ;kstuk
¼iv½ D;wfcd laxr ;kstuk
¼v½ foLrkfj f=dks.kh; laxr ;kstuk
¼vi½ o`Rrh; laxr ;kstuk
¼vii½ usLVsM xqzi ¼lewg½ foHkkT; laxr ;kstuk

mnkgj.k% eku yhft,] nks izkFkfed oa’kdze ¼1 ,oa 2½ vkSj 10 ¼3] 4] &&&& 12½ f}rh;d oa’kdze gSaA izR;sd izkFkfed oa’kdze dks vU; ‘ks”k oa’kdzeksa ds lkFk ladj.k fd;k tkrk gS ftlds ifj.kke 21 oa’kdzeksa ds lkFk ladj.k fd;k ftlds ifj.kke 21 oa’kdzeksa ds :i esa lkeus vkrk gSA f}rh;d oa’kdzeksa ds chp ladj.kksa ls f=dks.kh; laxr ;kstuk ij vk/kkfjr ,d ihMhlh dk fuekZ.k gksrk gSA ,d f=dks.kh; laxr ;kstuk esa 10 f}rh;d oa’kdzeksa dh ,d laHkkfor O;oLFkk fuEufyf[kr gS%

*	3	4	5	6
3	*	7	8	9
4	7	*	10	11
5	8	10	*	12
6	9	11	12	*

fofHkUu oa’kdzeksa ds izFke rFkk f}rh; ,lksfl,V esa mDr ,lksfl,’ku Ldhe dk vuqlj.k fd;k x;k tks fuEufyf[kr gS%

oa’kk oyh	1^{st} ,lksfl,V	2^{nd} ,lksfl,V
3	4, 5, 6, 7, 8, 9	10, 11, 12
4	3, 5, 6, 7, 10, 11	8, 9, 12

5	3, 4, 6, 8, 10, 12	7, 9, 11
6	3, 4, 5, 9, 11, 12	7, 8, 10
7	3, 8, 9, 4, 10, 11	5, 6, 12
8	3, 7, 9, 5, 10, 12	4, 6, 11
9	3, 7, 8, 6, 11, 12	4, 5, 10
10	4, 7, 11, 5, 8, 12	3, 6, 9
11	4, 7, 10, 6, 9, 12	3, 5, 8
12	5, 8, 10, 6, 9, 11	3, 4, 7

vc f}rh; oa'kkoyh ds lkFk blds izFke ,lksfl,V ds lHkh laHkkfor dzkl rS;kj fd, x, ftlesa 30 dzkWl izkIr gq,A bl izdkj vafre ,ihMhlh lyku esa 51 dzkWl dks uhps n'kkZ;k x;k gS tgka dzkWl dks x }kjk lanfHkZr fd;k x;k gS%

	oa'kdze												
oa'kdze		1	2	3	4	5	6	7	8	9	10	11	12
	1	-	x	x	x	x	x	x	x	x	x	x	x
	2		-	x	x	x	x	x	x	x	x	x	x
	3			-	x	x	x	x	x	x	x	x	x
	4				-	x	x	x	x	x	x	x	x
	5					-	x	x	x	x	x	x	x
	6						-	x	x	x	x	x	x
	7							-	x	x	x	x	x
	8								-	x	x	x	x
	9									-	x	x	x
	10										-	x	x
	11											-	x
	12												-

;fn mijksDr ,ihMhlh Iyku ds LFkku ij 12 isjsaVy ykbuksa ds lkFk ,d lhMhlh Iyku dk mi;ksx fd;k tkrk gS rks 66 dzkWlksa dh vko';drk gksxhA blh izdkj gesa nwljs ,lksfl,V dk iz;ksx djds vU; ,ihMhlh esfVax fMtkbu izkIr gksxk tks uhps fn;k x;k gS%

	oa'kdze											
oa'kdze		1	2	3	4	5	6	7	8	9	10	11	12
	1	-	x	x	x	x	x	x	x	x	x	x	x
	2		-	x	x	x	x	x	x	x	x	x	x
	3			-	x	x	x	x	x	x	x	x	x
	4				-	x	x	x	x	x	x	x	x
	5					-	x	x	x	x	x	x	x
	6						-	x	x	x	x	x	x
	7							-	x	x	x	x	x
	8								-	x	x	x	x
	9									-	x	x	x
	10										-	x	x
	11											-	x
	12												-

12 izeq[k fcanqvksa ds lkFk iw.kZ Mkb ,yhy ds fy, 66 dzkWlksa dh rqyuk esa dqy 36 dzkWl gSaA

baVjykbu rqyukvksa ds fofHkUu oxksZa ls lacaf/kr izlj.k vFkkZr V p ×p ¼nksuksa izkjafHkd oa'kdze½ V p ×p ¼,d izkjafHkd oa'kdze rFkk vU; nwljk oa'kdze½ Vq×q_c ¼f}rh; oa'kdzeksa esa rqyukvksa dk vkSlr izlj.k tks dzkl gqvk gS½ vkSj Vq×q_nc ¼f}rh; oa'kdzeksa esa rqyukvksa dk vkSlr izlj.k tks dzkl ugha gqvk gS½ bls ifjdfyr fd;k x;kA tc izFke ,lksfl,V dk bLrseky fd;k x;k rc ;g izlj.k 0.2000,0.2467,0.2963 rFkk 0.2593 vkSj tc f}rh; ,lksfl,V dk bLrseky fd;k x;k rc ;g 0.2000, 0.3614, 0.5556 rFkk 0.4444 FkkA

7- ijh{k.k dh rqyuk esa daVªksy ds fy, Mkb ,yhy dzkWl fMtkbu

;g ijh{k.kkRed fLFkfr gks ldrh gS tgka ijh{k.k ds vkjafHkd pj.k esa vusd u, oa'kdze fodflr gq, vkSj ;g vk'kk gS fd flQZ dqN oa'kdze gh vkxkeh fujh{k.k ds fy, csgrj ik, tk,axsA u, oa'kdzeksa dh igys ,d ¼;k vf/kd½ daVªksy oa'kdze ds lkFk rqyuk dh xbZ] ftls vkxkeh vUos"k.k ds fy, u, oa'kdze dh tkap esa ijh{k.kdrkZvksa }kjk mi;ksx fd;k tk,xkA ijh{k.kdrkZ dks

rqyukRed dk;Z djus ds fy, ijh{k.k rFkk daVªksy oa'kdze ds chp ;FkklaHko mRd`"Vrk dk;e j[kuh pkfg,A

pksbZ vkfn ¼2004½ us iw.kZ ;kn`fPNd fMtkbu ds fy, ekWMy ds rgr ijh{k.k oa'kdze ds lkFk daVªksy oa'kdze dh rqyuk ds fy, Mkb ,yhy dzkWl dk v/;;u fd;kA Vlw rFkk fjax ¼2005½ us CykWd fMtkbu ds fy, ekWMy ds rgr daVªksy oa'kdze ds lkFk nks ;k rhu ijh{k.k oa'kdze dh rqyuk ds fy, Mkb ,yhy dzkWl ijh{k.kksa dk ,d b"Vre v/;;u fd;k x;kA nkl vkfn ¼2006½ esa bl ij vkxkeh vUos"k.k fd;k x;k vkSj ,&vkIVhey fMtkbu ds fy, larks"ktud fLFkfr mRiUu dhA JhokLro ¼2010½ us ijh{k.k cuke fu;a=.k ¼VSLV oflZl daVªksy½ rqyukvksa ds fy, Mkb ,yhy dzkWl ijh{k.k ds fy, y?kq vkSj dq'kyre CykWd fMtkbu dh dqN QSfeyh l`tu dh fof/k miyC/k djkbZ gSA

VSLV oflZl daVªksy rqyukvksa ds fy, Mkb ,yhy dzkWl ijh{k.kksa gsrq izlj.k larqfyr CykWd fMtkbu% daVªksy oa'kdze ds lkFk t ijh{k.k oa'kdze dh rqyuk ds fy, Mkb,yhy dzkWl ijh{k.kksa ds fy, ,d CykWd fMtkbu izlj.k larqfyr gksxk ;fn lHkh ewyHkwr fo"kerk gca izHkko ls lacaf/kr gSA

i½ leku izlj.k ds lkFk vkdfyr ijh{k.k oa'kdzeksa ds chp $(Vt \times t'), (t \neq t')$

ii½ leku izlj.k $(V_{t\times c})$ ds lkFk vkdfyr ijh{k.k daVªksy oa'kdze ds chp

VSLV oflZl daVªksy rqyukvksa ds fy, Mkb ,yhy dzkWl ijh{k.kksa gsrq izlj.k larqfyr CykWd fMtkbu% daVªksy oa'kdze ds lkFk t ijh{k.k oa'kdze dh rqyuk ds fy, Mkb ,yhy dzkWl ijh{k.kksa ds fy, ,d CykWd fMtkbu izlj.k larqfyr gksxk ;fn lHkh ewyHkwr fo"kerk gca izHkko ls lacaf/kr gSA

i½ ijh{k.k oa'kdze tks i^{th} $(i = 1, 2, \ldots, m)$ gS tks ijLij ,lksfl,V gSa bldk leku:ih izlj.k $(Vt \times 'ii (tii \neq ti' = 1, 2, \ldots, ni$ tgka n_i i^{th} ,lksfl,V dh la[;k gS½ ds lkFk vkdyu fd;k x;k

ii½ ijh{k.k vkSj daVªksy oa'kdze dk vkdyu leku:ih izlj.k $¼V_{ti\times c}$ ½ ds lkFk vkdyu fd;k x;k tc rd ijh{k.k oa'kdze ijLij i^{th} ds lkFk lacaf/kr FksA

mnkgj.k% fuEufyf[kr Vªsaxqyj Ldhe ij fopkj fd;k ftls uhps n'kkZ;k x;k gS ;gka $t = 10$ ijh{k.k oa'kdze gSa%

*	1	2	3	4
1	*	5	6	7
2	5	*	8	9
3	6	8	*	10
4	7	9	10	*

daVªksy oa’kdze ds lkFk eq[; fod.kZ esa fjDr LFkku Hkjsa ¼0 }kjk n’kkZ;k x;k½A

0	1	2	3	4
1	0	5	6	7
2	5	0	8	9
3	6	8	0	10
4	7	9	10	0

leLr laHkkfor dzkWl dks izR;sd dkWye esa 10 vkdkj ds CykWd esa daVªksy oa’kdze ds lkFk 10 ijh{k.k oa’kdzeksa dh rqyuk ,d ihMhlh esfVax i;kZoj.kh; fMtkbu ls dh xbZA

CykWDl				
1	**2**	**3**	**4**	**5**
0×1	1×0	2×5	3×6	4×7
0×2	1×5	2×0	3×8	4×9
0×3	1×6	2×8	3×0	4×10
0×4	1×7	2×9	3×10	4×0
1×2	0×5	5×0	6×8	7×9
1×3	0×6	5×8	6×0	7×10
1×4	0×7	5×9	6×10	7×0
2×3	5×6	0×8	8×0	9×10
2×4	5×7	0×9	8×10	9×0
3×4	6×7	8×9	0×10	10×0

fMtkbu ds iSjkehVj gSa t = 10, c = 1, b = 5, r_{tc} = 2, k = 10, N_{total} = 50 vkSlr izlj.k gS Vtxt= 0.3968 rFkk Vt’×c = 0.2619

mnkgj.k% vkdkj 3 ds CykWd esa 5 oa’kdzeksa ds fy, fuEufyf[kr cxSj de dh gqbZ chvkbZch fMtkbu ij fopkj fd;k x;k%

CykWd

1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	2	2	2	3
2	2	2	3	3	4	3	3	4	4
3	4	5	4	5	5	4	5	5	5

izFke CykWd dks gVk;k x;k rFkk gVk, x, CykWd ds CykWd vkadM+ksa ij ijh{k.k oa'kdzeksa ds :i esa fopkj fd;k x;k ¼1]2]3½A 'ks"k oa'kkofy;ksa ¼4]5½ dks daVªksy oa'kkoyh ds :i esa ekuk x;k vkSj dze'k% 0_1 rFkk 0_2 ds :i esa n'kkZ;k x;kA 'ks"k CykWdksa ds rgr lHkh laHkkfor fof'k"V dzkWl cukus ds fy, geus 3 ijh{k.k oa'kdzeksa ds lkFk&lkFk daVªksy oa'kdzeksa dh rqyuk ds fy, fuEufyf[kr izlj.k larqfyr Mkb,yhy dzkWl esfVax i;kZoj.kh; fMtkbu izkIr fd,%

CykWd								
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1×2	1×2	1×3	1×3	$1\times0_1$	2×3	2×3	$2\times0_1$	$3\times0_1$
$1\times0_1$	$1\times0_2$	$1\times0_1$	$1\times0_2$	$1\times0_2$	$2\times0_1$	$2\times0_2$	$2\times0_2$	$3\times0_2$
$2\times0_1$	$2\times0_2$	$3\times0_1$	$3\times0_2$	$0_1\times0_2$	$3\times0_1$	$3\times0_2$	$0_1\times0_2$	$0_1\times0_2$

fMtkbu ds iSjkehVj gSa $t = 3,\ c = 2,\ b = 9,\ r_{tt} = 2,\ r_{tc} = 3,\ k = 3,\ N_{total} = 27$ rFkk vkdfyr izlj.k gSa

$V_{t\times t'} = 0.5000$ rFkk $V_{t\times c} = 0.4333$ A

,ebZvkjlh fMtkbu

;g ekurs gq, fd fof'k"V la;qDr n{krk ux.; gS blds rgr Mkb ,yhy@vkaf'kd Mkb ,yhy dzkWl ijh{k.kksa ij dk;Z fd;k x;kA lkekU; la;qDr n{krk dk vuqeku yxkus ds vykok] ijh{k.kdrkZ izk;% nksuksa thlh, izHkko rFkk ,llh, izHkko nksuksa ij vuqeku yxkus dk bPNqd gksrk gSA

pksbZ rFkk eq[kthZ ¼1999½ us Mkb,yhy dzkWl ijh{k.kksa ds b"Vre izHkko ij ml le; v/;;u fd;k tc fof'k"V la;qDr n{krkvksa dks Hkh ekWMy esa 'kkfey fd;k x;kA NksbZ ,V-vkWy- ¼2002½] nkl vkSj Ms ¼2004½ rFkk Ms ¼2010½ us lkekU; la;qDr n{krk vkdyu ds fy, yEcdks.kh; CykWd laiw.kZ Mkb,yhy dzkWl dh vuqdwyrk dks ml le; Li"V fd;k tc ekWMy esa Hkh fof'k"V la;qDr n{krk,a 'kkfey FkhA blds vykok] tc ijh{k.k lkexzh esa nks fn'kkvksa esa fotkrh;rk mi;ksxh gksaxsA bl izdkj dh fLFkfr esa esfVax i;kZoj.kh; iafDr] dkWye ¼,ebZvkjlh½ fMtkbu tgka dzkWl ,d iafDr dkye ls LFkkfir gS ogka bldk mi;ksx djuk ykHkdkjh gksrk gSA

lanHkZ

vxzoky ,l-lh- rFkk nkl ,e-,u- ¼1990½A vkaf'kd Mkb,yhys dzkWl ds fy, viw.kZ CykWd fMtkbuA la[;k ch 52] 75&81

vxzoky ,p-lh- ¼1985½] ,d pkj oxZ pdzh; ,lksfl,'ku Ldhe rFkk lacaf/kr vkaf'kd Mkb,yhy dzkWl] la[;k ch- 47] 78&90

vk;kZ ,-,l- ¼1983½A vkaf'kd Mkb,yhys dzkWl ds fy, o`Rrkdkj ikni] ck;kseSfVªd 39] 43&52

vk;kZ ,-,l- rFkk ukjk;.k ih- ¼1977½ rhu vkSj pkj ,lksfl,V oxksZa ds lkFk dqN ,lksfl,'ku Ldhe ij vk/kkfjr vkaf'kd Mkb,yhy dzkWl la[;k ch 39] 394&399

pksbZ ds-lh-] xqIrk ,l- rFkk dkxh;kek ,l- ¼2004½A ijh{k.k ofxZl daVªksy rqyuk ds fy, Mkb,yhys dzkWl gsrq fMtkbuA ;wfVfyVl eSFkeSfVdk] 65] 167&180

nkl ,l-,u- rFkk f'ko jke ds ¼1968½A vkaf'kd Mkb,yhys dzkWl rFkk viw.kZ CYkkWd fMtkbu MkW- ih th ikuls dk QSfyfl,'ku okY;we- bafM;u lkslk0 vkQ ,xzh0 LVSfVLV 49&59

nkl ,] xqIrk ,l- rFkk dkth;kek] ,l- ¼2006½A ijh{k.k oflZl daVªksy rqyuk ds fy, ,&vkIVhdy Mkb,yhy dzkWl ts- ,Iiyh- L?VsfVLV 33 ¼6½] 601&608

QkbQ] ts-,y- rFkk fxycVZ ,u- ¼1963½A vkaf'kd Mkb,yhys dzkWl ck;kseSfVªDl] 19] 278&286

?kks"k Mh-ds- rFkk fnfo/kk ts- ¼1997½A nks ,lksfl,V Dykl ihchvkbZch fMtkbu rFkk vkaf'kd Mkb,yhys dzkWl ck;kseSfVªdk] 84 ¼1½] 245&248

fxycVZ] ,u- ¼1958½A ikni iztuu esa Mkb,yhys dzkWl gSjhfMVh] 12] 477&492

fgdhyeku rFkk dSE;ksuZ Mh ¼1963½A oxZ foHkkT; vkaf'kd Mkb,yhys dzkWl ds nks oxZ ck;kseSfVªdk] 50 ¼3 rFkk 4½] 281&291

fgadhyeku] ds- rFkk LVuZ] ds- ¼1960½A dztquxlIysu twV lsyhdfVvksUl tqadVqax csbZ ckSeu- flYosbZ tsusV 9] 121&133

glq okbZ-,Q- rFkk fVaMk lh-ih- ¼2005½A ijh{k.k mipkj ds lkFk daVªksy dh rqyuk ds fy, ,- vkIVhey rFkk dq'ky Mkb,yhy dzkWl ijh{k.k LVSfVLV izksc- ySVlZ 71 ¼1½] 99&110

tXxh ,l- rFkk 'kqDyk vkj-ds- ¼1996½A laiw.kZ Mkb,yhy dzkl ds fy, laof/kZr vkaf'kd Mkb,yhy dzkWl dh rqyuk bafM;u t- tsusV- IykWV czhM 56 ¼3½] 341&349

dkSf'kd ,y-,l- rFkk iqjh ih-Mh- ¼1989½A lkekU;d`r mfpr dks.kh; ,lksfl,'ku Ldhe ij vk/kkfjr vkaf'kd Mkb,yhy dzkWl dE;wu- LVsfVLV% F;ksjh eSFk 18 ¼7½] 2501&2510

dkSf'kd ,y-,l- ¼1999½- rhu ,lksfl,V Dykl ,lksfl,'ku Ldhe ij vk/kkfjr vkaf'kd Mkb,yhy dzkWl ts- ,Iyh- LVsfVLV 26 ¼2½] 195&201

dSEiFkksuZ] vks- rFkk dwj.kksa] vkj-,u- ¼1961½A vkaf'kd Mkb,yhy dzkWl ck;kseSfVªDl 17] 229&250

dqfj;kdksl] ,l- ¼1998½A larqfyr rFkk vkaf'kd larqfyr v/kwjs CykWd fMtkbu ij v/;;u vizdkf'kr ih,pMh- Fkhfll] Hkk-d`-v-la-] ubZ fnYyh

ekFkqj ,l-,u- rFkk ukjk;.k ih- ¼1976½A vkaf'kd Mkb,yhy dzkl ds fy, dqN b"Vre ;kstuk,a bafM;u t- tsusV 36 ¼3½] 301&308

ukjk;.k ih- lqCckjko] lh- rFkk fuxe] ,-ds- ¼1974½ izlkfjr f=Hkqtkdkj ,lksfl,'ku Ldhe ij vk/kkfjr vkaf'kd Mkb,yhy dzkWl bafM;u tuZ0 tsusV 34] 309&317

ukjk;.k ih- rFkk vk;kZ ,-,l- ¼1981½A Vªadsf VM f=Hkqtkdkj ,lksfl,'ku Ldhe vkSj lacaf/kr vkaf'kd Mkb,yhys dzkl la[;k ch- 43 ¼1½] 93&103

iSMjlu Mh-th- ¼1980½A laof/kZRk vkaf'kd Mkb,yhy dzkl gSjhfMVh 44] 329&331

flag ,e- rFkk fgafdyeku ds- ¼1995½- viw.kZ CykWdksa esa vkaf'kd Mkb,yhy dzkWl ck;kseSfVªdk] 51] 1302&1314

'kekZ ,e-ds- ¼1998½A o`Rrkdkj fMtkbu }kjk vkaf'kd Mkbysy dzkWl ts- bafM;u lkslk-,xzh- LVsfVLV 51 ¼1½] 17&27

JhokLro ,l- ¼2010½A laof/kZr vkaf'kd Mkb,yhy dzkWl fMktbu ij dqN vUosd"k.k vizdkf'kr ,e,llh Fkhfll] Hkk-d`-v-la-] ubz fnYyh

ofxZl lh- tXxh] ,l- 'kekZ] oh-ds- rFkk flag ;w-oh- ¼2005½A vkaf'kd Mkb,yhy dzkl esa vkaf'kd larqfyr viw.kZ CykWd fMtkbu dk mi;ksxA bafM;u tuZ0 tsusV 65 ¼1½] 37&40

osadVje.k] vkj ¼1985½A izxqf.kr rjhds ds dzkl ds dqN igyw- ih,pMh Fkhfll] Hkk-d`-v-la-] ubZ fnYyh

डिटेक्शन एवं क्यू॰ टी॰ एल एस्टिमेशन

डॉ. हिमाद्रि रॉय
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

परिचय

क्वांटिटेटिव ट्रेट लोकी ;क्यूटीएलद्ध मैपिंग एक अध्ययन है और विभिन्न जीनोमिक स्थानों पर जीनोटाइप के बीच संबंध के बारे में निष्कर्ष निकालता है और मात्रात्मक लक्षणों के एक सेट के लिए फेनोटाइप जिसमें संख्याए जीनोमिक स्थितिए प्रभाव और क्यूटीएल की बातचीत शामिल है। तो, मात्रात्मक लक्षणों की आनुवंशिक संरचना बहुत जटिल है। हाल के वर्षों में, आनुवंशिक मार्कर मैप्स की बड़े पैमाने पर उपलब्धता के कारण मात्रात्मक विशेषता लोकी (क्यूटीएल) मैपिंग का उपयोग करके मात्रात्मक विशेषता की जटिल प्रकृति का अध्ययन करने में मदद मिलती है और क्योंकि जटिल आनुवंशिक वास्तुकला के आरेखण की कठिनाई के कारण, सबसे अधिक क्यूटीएल मैपिंग प्रयोगों को डिज़ाइन किए गए प्रयोगों के माध्यम से किया जाता है, जैसे बैकक्रॉस, F2 और पुनः संयोजक इनब्रेड लाइनें, जहां एक अलग-अलग स्थानों पर एलील्स की संख्या ज्ञात एलील आवृत्तियों के साथ दो तक सीमित है। आज तक, क्यूटीएल मैपिंग के लिए अलग-अलग सांख्यिकीय पद्धति का पता लगाया गया है और प्रभाव का आकलन किया गया है। सांख्यिकीय दृष्टिकोण से, क्यूटीएल मैपिंग के इन तरीकों को तीन व्यापक वर्गों में विभाजित किया गया है: प्रतिगमन, अधिकतम संभावना और बायेसियन मॉडल। क्यूटीएल मैपिंग का पहला और महत्वपूर्ण उद्देश्य यह परीक्षण करना है कि क्यूटीएल की उपस्थिति है या नहीं। क्यूटीएल की उपस्थिति का परीक्षण परीक्षण के आँकड़ों जैसे कि छात्र की टी-टेस्ट, दो सैंपल टी-टेस्ट, विश्लेषण का विश्लेषण (ANOVA) तकनीक आदि से किया जा सकता है। जब क्यूटीएल की उपस्थिति की पहचान की जाती है, तो हमारा अगला दृष्टिकोण मार्कर स्थानों यानी जीनोम मैपिंग, उनके आत्मविश्वास अंतराल और पहचाने गए क्यूटीएल के दोनों तरफ फ्लैंकिंग मार्करों की पहचान करना है।

क्यूटीएल मैपिंग के तरीके

क्यूटीएल का पता लगाने के लिए मुख्य रूप से दो विधियों का उपयोग किया जाता हैरू ;1द्ध एकल मार्कर विश्लेषण और ;2द्ध अंतराल मानचित्रण। अंतराल मानचित्रण विधि को भी दो प्रक्रिया से विभाजित किया जाता हैए ;पद्ध सरल अंतराल मानचित्रण और ;पपद्ध समग्र अंतराल मानचित्रण।

एकल मार्कर विश्लेषण

जटिल सांख्यिकीय विधियों में गहरी जाने से पहलेए एकल मार्कर विश्लेषण ;एसएमएद्ध क्यूटीएल मैपिंग ;एडवर्ड एट अल।ए 1987 और वेलर एट अल।ए 1988द्ध के तंत्र को समझने में मदद करता है। एकल.मार्कर विश्लेषण इस विचार पर आधारित है कि यदि एक मार्कर जीनोटाइप और विशेषता मान के बीच एक संबंध हैए तो संभावना है कि एक क्यूटीएल उस मार्कर स्थान के करीब है। इस पद्धति मेंए क्यूटीएल.मार्कर एसोसिएशन को खोजने के लिए एक समय में एक मार्कर शामिल होता है। आमतौर पर इस्तेमाल की जाने वाली सांख्यिकीय तकनीकों जैसे टी-टेस्ट, एनोवा, रैखिक प्रतिगमन, संभावना अनुपात परीक्षण और अधिकतम संभावना आकलन (हेली एंड नॉट, 1992; निएनहिस एट अल।, 1987; वांग एट अल, 1994) इस पद्धति का उपयोग किया जाता है।

क्यूटीएल प्रतिगमन मॉडल

प्रतिगमन तरीके आम तौर पर बहुत आसान होते हैं और तेजी से कम्प्यूटेशनल रूप से। अधिकतम संभावना कम्प्यूटेशनल रूप से अधिक मांग हैए और विशिष्ट सॉफ़्टवेयर की आवश्यकता है। कई डिजाइनों के लिएए परिणाम प्रतिगमन के समान हैं। यह प्रतिगमन विश्लेषण को आकर्षक बनाता है क्योंकि इसका उपयोग तरीकों को फिर से जमा करने में किया जा सकता है। हम कई मार्करों ;2 से अधिकद्ध और एक से अधिक क्यूटीएल के लिए खाते के तरीकों के साथ क्यूटीएल मैपिंग पर भी चर्चा करेंगे। अन्य QTL के लिए लेखांकन सहायक कारक सहित, या समग्र अंतराल मानचित्रण का उपयोग

करके प्रस्तावित किया गया है। वे मिश्रित मॉडल विधियां हैं और मोंटे कार्लो मार्कोव चेन विधियां हैं। दोनों विधियों में, क्यूटीएल को निर्धारित या यादृच्छिक प्रभावों के रूप में मॉडल किया जाता है, और अतिरिक्त यादृच्छिक प्रभाव पॉलीजेनिक भिन्नता के लिए जिम्मेदार हो सकते हैं। मार्कर डेटा से क्यूटीएल जीनोटाइप संभावनाओं का अनुमान लगाने के लिए संयुक्त अलगाव और लिंकेज विश्लेषण की आवश्यकता है। मिश्रित मॉडल विधियाँ युग्मक संबंध मैट्रिक्स पर आधारित हैं, जिन पर संक्षेप में चर्चा की जाएगी। एकल मार्कर या एकाधिक मार्कर का उपयोग करके एक मात्रात्मक विशेषता और आनुवंशिक मार्करों के बीच संबंध का मूल्यांकन किया जा सकता है। एक एकल मार्कर का उपयोग करते समय, उस मार्कर से जुड़े क्यूटीएल के अलगाव के बारे में अनुमान लगाना संभव है। हालांकि, एकल मार्करों के मामले में, क्यूटीएल प्रभाव के आकार और उसकी स्थिति (मार्कर के सापेक्ष) के बीच अंतर करना संभव नहीं है। इसके अलावा, एकल मार्कर विश्लेषण में कम शक्ति होती है अगर मार्कर बहुत दूर हैं। यदि दो (या अधिक) मार्करों को संयुक्त रूप से एक विश्लेषण में उपयोग किया जाता है, तो क्यूटीएल प्रभाव की स्थिति और आकार के बीच बहुत कम उलझाव होता है, और क्यूटीएल का पता लगाने में अधिक शक्ति होती है, भले ही मार्कर दूर हो। क्यूटीएल प्रभाव के बारे में और साथ ही क्यूटीएल और मार्कर (यानी क्यूटीएल की स्थिति) के बीच पुनर्संयोजन दर के बारे में अनुमान संभव है। मार्करों के बीच पुनर्संयोजन दर आमतौर पर ज्ञात है। इसलिए एक क्यूटीएल की मैपिंग के लिए विश्लेषकों में कई मार्कर जीनोटाइप के उपयोग की आवश्यकता होती है। क्यूटीएल मैपिंग के लिए अग्रिम सांख्यिकीय पद्धति पर चर्चा करने से पहले, हमें चूहों के लिए एक बैकक्रॉस डिज़ाइन मान लेना चाहिए। एक बैकक्रॉस डिज़ाइन में जीनोटाइप्ड मार्कर और माउस बॉडी वेट के लिए एक डेटा संरचना
तालिका 1रू माउस डेटा संरचना

Sample	**Marker**		**Body weight**
	A	**B**	
1	1	1	30
2	1	1	32
3	1	1	28
4	1	1	29
5	1	0	29
6	0	1	22
7	0	0	20
8	0	0	21
9	0	0	20
10	0	0	21

इस उदाहरण में शरीर के वजनए लए और दो ज्ञात जीनोटाइप्स के साथ एकुजस के लिए फेनोटाइप्ड दस चूहों शामिल हैंए फु ;1 द्वारा इंगितद्ध और ु ;0 द्वारा इंगितद्ध। ऐसा प्रतीत होता है कि क्यूटीएल जीनोटाइप क्यूक्यू को ले जाने वाले चूहों की तुलना में अधिक भारी होता हैए क्योंकि वे जीनोटाइप क्यूईके ले जाते हैंए हालांकि जो चूहों में एसएई जीनोटाइप होता हैए वे शरीर के वजन के बराबर नहीं होते हैं। परीक्षण करने के लिएए यह वास्तव में मामला है और शरीर के वजन पर क्यूटीएल के प्रभाव का अनुमान लगाता हैए हम एक साधारण प्रतिगमन मॉडल के रूप में मानते हैं

$$y_i = \mu + \alpha x_i + e \qquad (1)$$

जहांए ल मनाया गया फेनोटाइप हैए समग्र अर्थ हैए योज्य प्रभाव हैए गप सूचक चर है जो माउस प के क्यूटीएल जीनोटाइप को निर्दिष्ट करता है और इसे परिभाषित किया जाता है।

$$x_i = \begin{cases} 1 \text{ if QTL genotype is qq} \\ 0 \text{ if QTL genotype is qq} \end{cases}$$

मप यादृच्छिक त्रुटि को आम तौर पर सामान्य रूप से वितरित माना जाता है। इस उदाहरण का उपयोग करते हुएए हम ल को शरीर के वजनए ग को मार्कर और मॉडल के रूप में पाते हैं और इसे दिया जाता है।

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

मापदंडों के कम से कम वर्ग अनुमान हैं

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

यहाँए यह पाया जाता है और। इन अनुमानों का उपयोग करकेए वर्गों ;एसएसटीद्ध की कुल राशि के रूप में दिया गया है

$$\sum_{i=1}^{10}(y_i - \mu)^2 = 205.6$$

और वर्गों का अवशिष्ट योग ;त्रैद्ध के रूप में दिया जाता है

$$\sum_{i=1}^{10}(y_i - \mu - x_i a)^2 = 12$$

क्यूटीएल आनुवंशिक प्रभाव ;एद्ध का महत्व तब एफ.मूल्य की गणना करके परीक्षण किया जाता है

$$F = \frac{(SST - SSE)/(2-1)}{SSE/(10-2)} = 129.07$$

अंत में महत्वपूर्ण मूल्य के साथ तुलना मेंए हम यह निष्कर्ष निकालते हैं कि यह क्यूटीएल बैकक्रॉस में शरीर के वजन पर एक महत्वपूर्ण प्रभाव डालती है।

दो सैंपल टी.टेस्ट

तालिका 1 में माउस बैकक्रॉस डेटा दो लिंक किए गए आणविक मार्कर ए और बी के लिए जीनोटाइप किए गए हैं और तालिका 1 में दिए गए हैं। प्रत्येक मार्कर पर दो जीनोटाइप 1 और 0ण द्वारा निरूपित किए जाते हैं। इन दोनों मार्करों के बीच का जुड़ाव उनकी स्थिरता से देखा जा सकता है। नमूने के बीच जीनोटाइपए चूहों 5 और 6 को छोड़कर। दो मार्करों के बीच पुनर्संयोजन अंश आर त्र 2ध्10 त्र 0ण्2 है। हम इन दो मार्करों का अलग.अलग विश्लेषण करेंगे। मार्कर ए को देखते हुए, ऐसा लगता है कि मार्कर जीनोटाइप के दो समूह शरीर के वजन में भिन्न हैं। एक प्रश्न स्वाभाविक रूप से उठता है कि क्या मार्कर ए में जीनोटाइप 1 और 0 के बीच शरीर के वजन में यह अंतर सांख्यिकीय रूप से महत्वपूर्ण है। यह एक दो-नमूना टी परीक्षण द्वारा परीक्षण किया जा सकता है। आज्ञा देना μ1 और μ0 क्रमशः जीनोटाइप 1 और 0 के साथ चूहों के दो अलग-अलग समूहों का वास्तविक विशेषता साधन है, और m1 और m0 को इसी नमूना साधन होने दें। परीक्षण के लिए परिकल्पना के रूप में तैयार किया जा सकता है

$$H_0 : \mu_1 = \mu_0$$

$$\boldsymbol{H_0 : \mu_1 \neq \mu_0}$$

टी टेस्ट स्टेटिस्टिक का उपयोग दो साधनों के बीच अंतर के महत्व के लिए किया जाता है

$$t = \frac{m_1 - m_0}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}$$

जहाँ ` 2 द्वारा दिया गया नमूना नमूना विचरण है

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}$$

द1ए द0 और के साथए क्रमशः दो अलग.अलग मार्कर समूहों में नमूना आकार और संस्करण हैं। शून्य परिकल्पना भ0 को अस्वीकार कर दिया जाएगा यदि टी परीक्षण सांख्यिकीय गणना टी.वितरण से प्राप्त किए जाने वाले महत्वपूर्ण मूल्य से अधिक या बराबर है। अगर हम ऊपरी α महत्वपूर्ण बिंदु को ज ;αए νद्ध से निरूपित करते हैंए तो हम α त्र 0ण05 पर परिकल्पना को अस्वीकार करते हैं यदि जझ ज ;0ण025ए νद्धए 0ण05 महत्व स्तर के लिए दो.पूंछ वाले टी मानए ν त्र द1 के साथ द0 . स्वतंत्रता की 2 डिग्री।

उदाहरणरू तालिका 1ए द1 त्र द0 त्र 5 के साथ दो मार्कर वर्ग देती हैए हम उ1 त्र 29ण6 और उ0 त्र 20ण8ए `1 त्र 0ण8367ए `2 त्र 1ण5166 और ` त्र 1ण50 के साथ मार्कर की गणना करते हैं। मान त्र 11ण3608 और महत्वपूर्ण मान। हम यह निष्कर्ष निकाल सकते हैं कि मार्कर ए शरीर के वजन के साथ महत्वपूर्ण रूप से जुड़ा हुआ है।

विश्लेषण का विश्लेषण ;एनोवाद्ध

थ्2 आबादी के लिएए मार्कर जीनोटाइप के तीन अलग.अलग समूह हैंए जिन्हें क्रमशः प्रत्येक मार्कर पर 2ए 1ए और 0 द्वारा दर्शाया जा सकता है ;तालिका 1 देखेंद्ध। तीन जीनोटाइप के बीच समग्र अंतर का परीक्षण करने के लिएए विचरण ;एनोवाद्ध के पारंपरिक विश्लेषण का उपयोग किया जा सकता है। तीन मार्कर जीनोटाइप्स के बीच अंतर के कारण माध्य वर्ग उस डिग्री को दर्शाता है जिसमें मार्कर एक विशेष गुण के लिए एक पुटीय क्यूटीएल से जुड़ा होता हैए जबकि मीन वर्ग जीनोट के भीतर अंतर के कारण होता है।

गणना की गई थ् मूल्य की तुलना थ् वितरणए थ्0ण05ए ;2ए दद्ध 3द्ध से प्राप्त महत्वपूर्ण मूल्य से की जाती है। एक महत्वपूर्ण मार्कर के कारण आनुवंशिक भिन्नता का अनुमान अनुमानित वर्ग ;तालिका 2द्ध से मतलब वर्ग ;एमएसद्ध के समीकरण और परिणामी समीकरण को हल करके किया जा सकता हैरू

$$\sigma_g^2 = \frac{MS_1 - MS_2}{k}$$

मार्कर द्वारा बताई गई मात्रात्मक विशेषता में फेनोटाइपिक विचरण का अनुपातए व्यापक बोधगम्यताए द्वारा अनुमानित है

$$\sigma_g^2 = \frac{MS_1 - MS_2}{k} \qquad (2)$$

फेनोटाइपिक भिन्नता के लिए ए मार्कर के योगदान का आकलन करने के लिए एक पैरामीटर के रूप में इस अनुपात का व्यापक रूप से उपयोग किया जाता है।

तालिका 2रू थ्2 आबादी में तीन जीनोटाइप समूहों के बीच अंतर के लिए एनोवा का सारांश

Source of Variation	Df	Mean Square	*F*-value	Expected Mean Square
Among marker genotypes	2	MS_1	MS_1/MS_2	$\sigma_e^2 + k\sigma_q^2$
Within marker genotypes	$n-3$	MS_2		σ_e^2

थ्2 आबादी में तीन मार्कर जीनोटाइप के बीच समग्र अंतर या तो ंकपजपअम या प्रभुत्व प्रभावए या दोनों के कारण हो सकता है। टी परीक्षण का उपयोग करके इन दोनों प्रभावों के महत्व को भी परखा जा सकता है। मार्कर के योगात्मक प्रभाव का परीक्षण करने के लिएए हमारे पास परीक्षण आँकड़ा है

$$t_1 = \frac{m_2 - m_0}{\sqrt{s^2\left(\frac{1}{n_2} + \frac{1}{n_0}\right)}} \qquad (3)$$

साथ से

$$s^2 = \frac{(n_2 - 1)s_2^2 + (n_0 - 1)s_0^2}{n_2 + n_0 - 2}$$

और मार्कर के प्रभुत्व प्रभाव का परीक्षण करने के लिएए हमारे पास परीक्षण आँकड़ा है

$$t_2 = \frac{m_1 - \frac{1}{2}(m_2 + m_0)}{\sqrt{s^2\left(\frac{1}{4n_2} + \frac{1}{n_1} + \frac{1}{4n_0}\right)}} \qquad (4)$$

$$s^2 = \frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_2 + n_1 + n_0 - 3}$$

जहांए और क्रमशः थ्2 के तीन अलग.अलग मार्कर समूहों में नमूना संस्करण हैं।

उदाहरणरू

तालिका 1 दस चूहों के साथ थ्2 की आबादी के लिए एक उदाहरण प्रदान करती हैए प्रत्येक को शरीर के वजन के लिए मापा जाता है और दो कोडिनेटर मार्कर ए और बी के लिए जीनोटाइप किया जाता है। हम मार्करों के लिए एमएस 1 त्र 65ण्47 और एमएसपी त्र 10ण्67 के बीच अंतर वर्गों के लिए माध्य वर्गों की गणना करते हैं। एए जिसमें से एफ.मूल्य 6ण्14 के रूप में गणना की जाती है। महत्वपूर्ण थ्0ण्05ए ;2ए10 त्र3 त्र 7द्ध त्र 4ण्7374 मान के साथ तुलना मेंए यह मार्कर शरीर के वजन के साथ महत्वपूर्ण रूप से जुड़ा हुआ माना जाता है। इस मार्कर के कारण आनुवंशिक भिन्नता की गणना त्र 5 के रूप में की जाती है।

मार्कर ए के लिएए तीन जीनोटाइप समूहों में द2 त्र 3ए द1 त्र 4ए और द0 त्र 3 होते हैंए और तीन नमूना साधनों की गणना उ2 त्र 30ए उ1 त्र 25ए और उ0 त्र 20ण्67 और तीन नमूने अंतपंद के रूप में गणना की जाती है त्र 4ए। त्र 22ए और त्र 0ण्3333ए क्रमशः।

हम क्रमशः ;3द्ध और ;4द्ध के साथ ज1 त्र 7ण्7658 और ज2 त्र –0ण्1220 के रूप मेंए ंककपजपअम और प्रमुख प्रभावों के लिए टी परीक्षण के आँकड़ों की गणना करते हैं। महत्वपूर्ण मानों के साथ तुलना में ज ;0ण्025ए ν त्र 3 3 त्र2 त्र 4द्ध त्र 2ण्1318 योज्य परीक्षण के लिए और ज ;0ण्025ए ν त्र 3 4 3–3 त्र 7द्ध त्र 1ण्8946 प्रभुत्व प्रभाव के लिएए हम उस मार्कर का निष्कर्ष निकालते हैं एक महत्वपूर्ण योगात्मक प्रभाव प्रदर्शित करता हैए लेकिन शरीर के वजन पर एक महत्वहीन प्रभुत्व प्रभाव।

मार्कर बी के लिए एक समान कंप्यूटिंग प्रक्रिया ली गई है। इस मार्कर का एफ.मान 0ण्85 हैए यह सुझाव देता है कि इसका चूहों में शरीर के वजन के साथ कोई महत्वपूर्ण संबंध नहीं है। योज्य और प्रभुत्व प्रभावों के परीक्षण के लिए टी.मानों की गणना क्रमशः ज1 त्र 1ण्2264 और ज2 त्र 0ण्5635 के रूप में की जाती है। यह देखा जा सकता है कि अतिरिक्त प्रभाव और प्रभुत्व प्रभाव दोनों ही गैर.महत्वपूर्ण हैं।

उदाहरण

हम सुगियामा एट अल के आंकड़ों पर विचार करेंगे। ;फिजियोल। जीनोमिक्स 10रू 5.12ए 2002द्ध। डेटा ठ ।स्ठ ध बश्र और ब्ठ। ध ब्श्र के बीच एक इंटरक्रॉस से हैंय केवल पुरुष संतानों को माना जाता था। चार फेनोटाइप हैंरू रक्तचापए हृदय गतिए शरीर का वजन और हृदय का वजन। हम ब्लड प्रेशर फेनोटाइप पर ध्यान केंद्रित करेंगेए जीनोटाइपेडेटा के साथ सिर्फ 163 व्यक्तियों पर विचार करेंगे और सादगी के लिए ऑटोसोम्स पर ध्यान केंद्रित करेंगे। डेटा अल्पविराम.सीमांकित फ़ाइल

ीजजचरूध्ध्णतुजसण्वतहध्नहण्बे में निहित हैं

सारांशरू डेटा ऑब्जेक्ट चीनी जटिल हैय इसमें जीनोटाइप डेटाए फेनोटाइप डेटा और जेनेटिक मैप शामिल हैं। आर के पास एक निश्चित राशि ंऑब्जेक्ट ओरिएंटेडष सुविधाएं हैंए ताकि सारांश और प्लॉट जैसे कार्यों के लिए कॉल की व्याख्या ऑब्जेक्ट के लिए उचित रूप से की जाती है।

विभिन्न भूखंडों के साथ आंकड़ों का सारांश यहां दिया गया है.

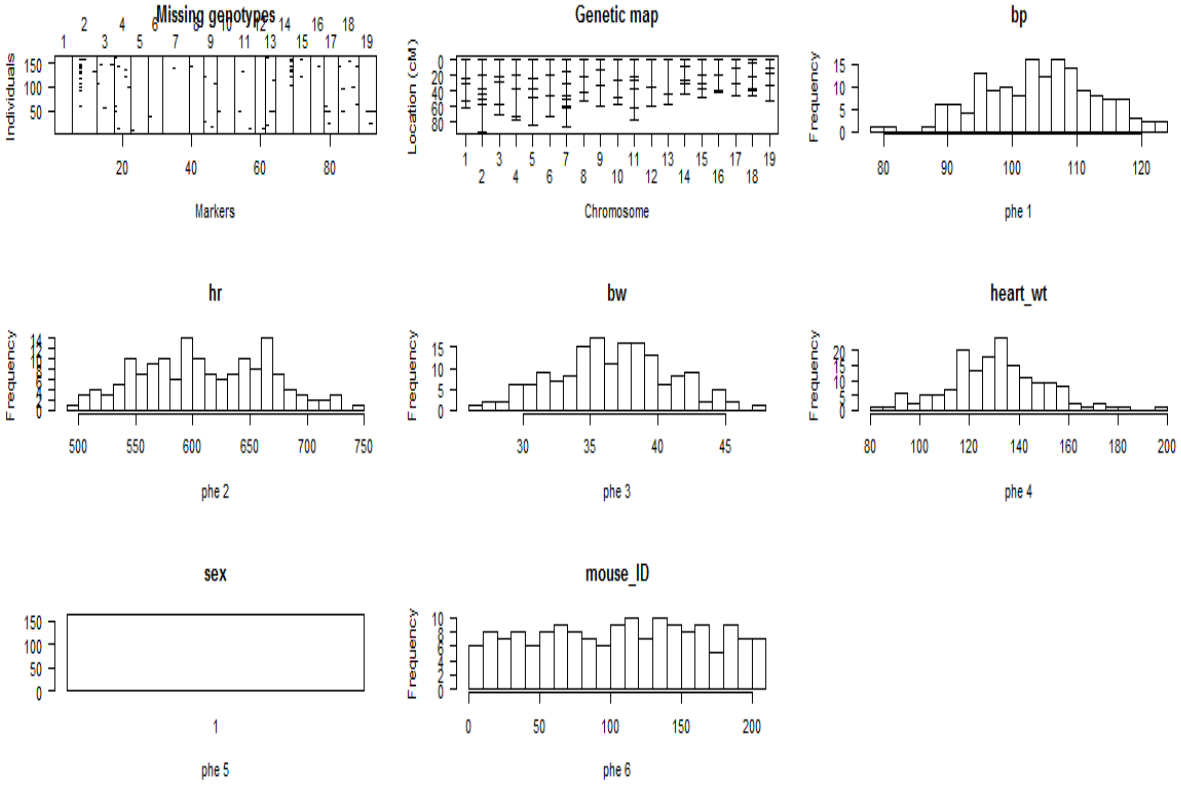
```
F2 intercross

No. individuals:    163

No. phenotypes:     6
Percent phenotyped: 95.1 95.7 99.4 99.4 100 100

No. chromosomes:    19
    Autosomes:      1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Total markers:      93
No. markers:        5 7 5 5 5 4 8 4 4 5 6 3 3 5 5 4 4 6 5
Percent genotyped:  98.3
Genotypes (%):      CC:23.9  CB:50.2  BB:26.0  not BB:0.0  not CC:0.0
```



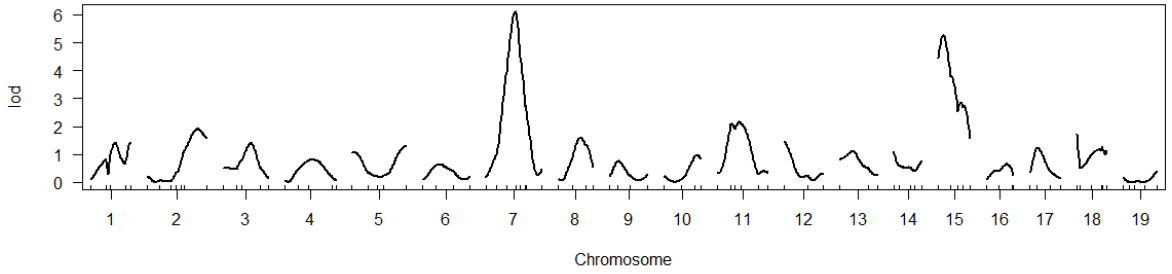
अंतराल मानचित्रण द्वारा एकल.क्यूटीएल विश्लेषण

आइए अब सिंगल.क्यूटीएल मॉडल के माध्यम से क्यूटीएल मैपिंग के लिए आगे बढ़ें। हम पहले क्यूटीएल जीनोटाइप संभावनाओं की गणना करते हैंए मनाया मार्कर डेटा दिया जाता है। यह मार्करों पर और क्रोमोसोम के साथ एक ग्रिड पर किया जाता है। तर्क चरण ग्रिड का घनत्व ;सीएम मेंद्ध हैए और बाद में क्यूटीएल विश्लेषण के घनत्व को परिभाषित करता है।

तालिका 6रू संबंधित स्थिति और स्व स्कोर के साथ मार्कर डेटा

Marker	chr	pos	LOD

D1MIT36	1	76.73	1.449
c2.loc77	2	82.80	1.901
c3.loc42	3	52.82	1.393
c4.loc43	4	47.23	0.795
D5MIT223	5	86.57	1.312
c6.loc26	6	27.81	0.638
c7.loc45	7	47.71	6.109
c8.loc34	8	54.40	1.598
D9MIT71	9	27.07	0.769
c10.loc51	10	60.75	0.959
c11.loc34	11	38.70	2.157
D12MIT145	12	2.23	1.472
c13.loc20	13	27.26	1.119
D14MIT138	14	12..52	1.119
c15.loc8	15	11.52	5.257
c16.loc31	16	45.09	0.647
D17MIT16	17	17.98	1.241
D18MIT22	18	13.41	1.739
D19MIT71	19	56.20	0.402



अंजीरू लोद बनाम क्रोमोसोम

यहां हमने गुणसूत्र 7 और 15 पर क्यूटीएल के लिए अच्छे सबूत देखे हैं

क्यूटीएल स्थान का अंतराल अनुमान

क्यूटीएल के स्थान का अंतराल अनुमान आमतौर पर 1.5-एलओडी समर्थन अंतराल के माध्यम से प्राप्त किया जाता है, जिसे फ़ंक्शन लॉंट के माध्यम से गणना की जा सकती है। वैकल्पिक रूप से, एक अनुमानित बेय्स विश्वसनीय अंतराल प्राप्त किया जा सकता है।

पहली और आखिरी पंक्तियाँ अंतराल के छोर को परिभाषित करती हैं; मध्य पंक्ति अनुमानित क्यूटीएल स्थान है:

पहली और आखिरी पंक्तियाँ अंतराल के छोर को परिभाषित करती हैंय मध्य पंक्ति अनुमानित क्यूटीएल स्थान हैरू

LOD interval	Bayes Interval

	Chr	Position	Lod		Chr	Position	Lod
c7.loc34	7	36.71	4.40	c7.loc37	7	39.71	5.09
c7.loc45	7	47.71	6.11	c7.loc45	7	47.71	6.11
c7.loc54	7	56.71	4.51	c7.loc50	7	52.71	5.38

निकटतम फ्लैंकिंग मार्करों की पहचान करें

LOD interval				Bayes Interval			
	Chr	Position	Lod		Chr	Position	Lod
D7MIT176	7	34.48	3.89	D7MIT176	7	34.48	3.89
c7.loc45	7	47.71	6.11	c7.loc45	7	47.71	6.11
D7MIT7	7	63.14	2.80	D7MIT323	7	54.45	4.69

जीनोमिक सिलेक्शन एवं प्रीडिक्शन

श्री उपेन्द्र प्रधान

भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

परिचय

प्राकृतिक विविधता दुनिया भर में पौधों की प्रजातियों के भीतर फेनोटाइपिक और आनुवंशिक विविधता का एक मूल्यवान और स्थायी संसाधन है जो पौधे के प्रजनन के लिए लाभदायक लक्षण प्रदान करते हैं। सहज, प्राकृतिक आनुवंशिक उत्परिवर्तन के कारण प्रजाति के भीतर की प्ररूपी भिन्नता जो विकासवादी, कृत्रिम और प्राकृतिक चयन प्रक्रियाओं द्वारा प्रकृति में बनी रहती है प्राकृतिक भिन्नता ने फसल आकृति विज्ञान और जैविक और अजैविक तनावों के प्रति उनकी प्रतिक्रिया को समझने के लिए बहुत प्रगति की वर्चस्व के लिए हजारों वर्षों के माध्यम से फसल पौधों में प्राकृतिक भिन्नता की समझ विकासात्मक लक्षणों और अनुकूली विशेषताओं के आनुवंशिक संशोधन में देखी जा सकती है जंगली प्रजातियों में प्राकृतिक भिन्नता अध्ययन ने घरेलू पौधों के अनुकूलन से संबंधित फेनोटाइपिक अंतर के आणविक आधार को स्पष्ट किया जो कि फेनोटाइपिक भिन्नता के रखरखाव और विकासवादी महत्व की व्याख्या करने के लिए महत्वपूर्ण है उदाहरण के लिए, जौ में छह-पंक्ति वाली 1 (वीआरएस 1) और गैर-भंगुर राचिस 1 (बीटीआर 1) या गैर-भंगुर राचिस 2 (बीआरआर 2) जीन पर हावी होने के परिणामस्वरूप स्पाइक आर्किटेक्चर फेनोटाइप पर स्पष्ट प्रभाव पड़ता है। वर्चस्व के दौरान, वीआरएस 1 जीन में फ़ंक्शन की हानि ने दो-पंक्ति वाली जौ को छह-पंक्ति में बदल दिया, जिससे अनाज की संख्या प्रति स्पाइक में बढ़ गई और बीटीआर जीन में विलोप गैर-भंगुर राचिस बनाते हैं जो अनाज प्रतिधारण में सुधार करते हैं जंगली और / या घरेलू, संवर्धित पौधों की विविधता के भीतर प्राकृतिक भिन्नता का विश्लेषण, फसल में सुधार के लिए विविध संसाधनों का कुशलतापूर्वक उपयोग करने में मदद कर सकता है और खेती किए गए फसल सुधार के आनुवंशिक आधार के ज्ञान में सुधार कर सकता है फसल पौधों में प्राकृतिक मात्रात्मक भिन्नता के आनुवंशिक विश्लेषण कुछ दशकों पहले विकसित किए गए थे एक जीन बैंक आनुवांशिक भिन्नता का एक समृद्ध स्रोत प्रदान करता है जो कि अनाज की पैदावार बढ़ाने और अजैविक और जैविक तनावों के प्रति सहिष्णुता में सुधार के लिए वांछित एलील को प्रजनन कार्यक्रमों में शामिल करने के माध्यम से खेती में सुधार करने के लिए उपयोग किया गया था। वर्चस्व और आधुनिक प्रजनन प्रक्रियाओं के दौरान होने वाली आनुवांशिक अड़चनें खेती में आनुवंशिक भिन्नता को कम करती हैं जो उत्पादकता, अनुकूलन और उपज स्थिरता को नकारात्मक रूप से प्रभावित करती हैं। डीएनए अनुक्रमण में हालिया प्रगति ने महत्वपूर्ण लक्षणों (अनाज की गुणवत्ता, जैविक और जैविक तनाव सहिष्णुता, आदि) को आनुवंशिक रूप से सुधारने का मार्ग प्रशस्त किया। अगली पीढ़ी के अनुक्रमण (एनजीएस) उदा। जीनोटाइपिंग-बाय-सीक्वेंसिंग (जीबीएस) हजारों एकल न्यूक्लियोटाइड बहुरूपता (एसएनपी) प्रदान करते हैं जो जौ गुणसूत्रों में सबसे जीनोमिक क्षेत्र को कवर करते हैं। लक्ष्य लक्षणों को नियंत्रित करने वाले एलील्स की पहचान करने के लिए कई शक्तिशाली सांख्यिकीय आनुवंशिकी विधियों का प्रस्ताव किया गया था। जीनोम-वाइड एसोसिएशन स्टडी (GWAS) उन उपयोगी तरीकों में से एक है और इसका उपयोग सफलतापूर्वक जौ में कई महत्वपूर्ण लक्षणों के लिए उम्मीदवार जीन की पहचान करने के लिए किया जाता है क्योंकि यह मार्कर प्रकार (जैसे एसएनपी) और एक लक्ष्य विशेषता के फेनोटाइप के बीच सहयोग का परीक्षण करता है। कई विचार और सिफारिशें हैं, जिन्हें ध्यान में रखा जाना चाहिए जब आनुवंशिकीविद् जीडब्ल्यूएएस प्रदर्शन करने का निर्णय लेते हैं। वर्तमान मसौदे में, हम GWAS के फायदे और दूरी, GWAS के प्रदर्शन के लिए विभिन्न तरीकों और GWAS परिणामों की व्याख्या करने के बारे में एक संक्षिप्त मार्गदर्शिका पर चर्चा करेंगे।

जटिल लक्षणों के आनुवंशिक अध्ययन

फॉरवर्ड जेनेटिक्स का उद्देश्य कई व्यक्तियों के फेनोटाइप को स्क्रीन करना है जो जीनोटाइपिक रूप से अलग हैं आनुवंशिक बहुरूपता और व्यक्तियों के बीच मनाया जाने वाला फेनोटाइपिक भिन्नता के बीच संबंधों को समझना मूलभूत हितों में से एक है इस बुनियादी संबंध का बड़े पैमाने पर अध्ययन किया गया है क्योंकि मेंडेल ने प्रदर्शित किया कि यह संबंध विरासत में मिला है। आनुवांशिक रूप से महत्वपूर्ण लक्षणों जैसे आनुवांशिक रूप से महत्वपूर्ण लक्षणों जैसे आनुवांशिक कारकों का खुलासा करने से एक विशिष्ट स्थान स्तर पर एलील भिन्नता की समझ की आवश्यकता होती है जो फेनोटाइप को नियंत्रित करती है और किसी दिए गए लक्षण की आनुवंशिक संरचना को नियंत्रित करती है। संयंत्र फेनोटाइप की भिन्नता सीधे मैपिंग दृष्टिकोणों का उपयोग करके अंतर्निहित करणीय लोकी से जुड़ी हुई है। इस लक्ष्य को प्राप्त करने के लिए, व्यक्तियों के बीच फेनोटाइपिक और जीनोटाइपिक अंतरों का अध्ययन या तो द्वि-अभिभावकीय क्यूटीएल मैपिंग आबादी (लिंकेज मैपिंग) या असंबद्ध व्यक्तियों की एसोसिएशन मैपिंग आबादी (एलडी मैपिंग) का उपयोग करके किया जाता है। इसलिए, दोनों मैपिंग दृष्टिकोण का उद्देश्य आणविक मार्करों की पहचान करना है जो क्यूटीएल से जुड़े हैं।

ये दृष्टिकोण आकर्षक और उपयोगी हो गए क्योंकि वे कई फसलों के लिए जीनोम अनुक्रमण और उच्च गुणवत्ता और घनत्व एसएनपी सरणियों में अग्रिमों का उपयोग करते हैं। एनजीएस जीबीएस का उपयोग करके आबादी को तेजी से मैप करने के लिए एक प्रभावी और अपेक्षाकृत कम लागत वाला दृष्टिकोण प्रदान करता है। जीबीएस तकनीक डीएनए नमूनों की जटिलता को कम करने के लिए प्रतिबंध एंजाइमों का उपयोग करती है और फिर उच्च गुणवत्ता वाले बहुरूपता डेटा का उत्पादन करती है भले ही एसएनपी का उपयोग करने वाले जीनोटाइपिंग बेहद कुशल और विश्वसनीय हैं, पिछले एक दशक में प्रदर्शन किए गए जीडब्ल्यूएएस ने कुछ कमियों का पता लगाया, जिन पर विचार किया जाना चाहिए। एसएनपी पर आधारित जीडब्ल्यूएएस पहले से मौजूद जेनेटिक वैरिएंट रेफरेंस पर निर्भर करता है जो व्यक्तियों को सीक्वेंसिंग और मैपिंग के लिए इस्तेमाल किया जाता है इस तरह के विशिष्ट डिजाइन से गायब पिनपॉइंट कारण परिवर्तनशील हो जाते हैं और अधिकांश आनुवंशिक संकेतों या जटिल लक्षणों के दुर्लभ उत्परिवर्तन का पता नहीं लगा सकते हैं

क्यूटीएल मैपिंग (लिंकेज मैपिंग)

लिंकेज या क्यूटीएल मैपिंग दृष्टिकोण आमतौर पर लक्ष्य लक्षणों को नियंत्रित करने वाले जीनोमिक क्षेत्रों (क्यूटीएल) की पहचान करने के लिए उपयोग किया जाता है। परिवार आधारित मानचित्रण विश्लेषण आनुवंशिक पुनर्संयोजन और अलगाव पर निर्भर करता है जो जैव-अभिभावकीय क्रॉस की संतानों में मानचित्रण आबादी के निर्माण के दौरान होता है, जिसके परिणामस्वरूप आनुवंशिक मानचित्रण संकल्प और एलील समृद्धि को प्रभावित करते हैं। क्यूटीएल मैपिंग साबित हुआ है और लोकी की पहचान करने के लिए एक शक्तिशाली दृष्टिकोण बना हुआ है जो अनुसंधान आबादी में रुचि के लक्षण के साथ सह-अलग हो जाता है। यह दृष्टिकोण विभिन्न प्रकार की आबादी में लागू किया जा सकता है उदा। एफ 2 **आबादी, डबल-अगुणित (डीएच) आबादी, बैकक्रॉस या पुनः संयोजक इनब्रेड लाइन्स (आरआईएल) परिवारों, प्रतिबंध खंड लंबाई बहुरूपता** (RFLP) **का उपयोग, प्रवर्धन टुकड़ा लंबाई बहुरूपता (**AFLP), **माइक्रोसैटेलाइट या सरल अनुक्रम दोहराने (**SSR) **और** SNP **चिह्नक। क्यूटीएल मैपिंग में प्रमुख मौलिक सीमाएं हैं कि माता-पिता के बीच एलीगेट को अलग करने की विविधता का केवल परीक्षण किया जा सकता है, और मैपिंग रिज़ॉल्यूशन केवल जनसंख्या विकास के दौरान होने वाली पुनर्संयोजन घटनाओं की संख्या पर निर्भर करता है। मैपिंग आबादी के लिए शुद्ध लाइनों (समरूप लाइनों) का विकास करने में समय लगता है और कम क्यूले रिचता के साथ एक संकीर्ण आधार में जिसके परिणामस्वरूप जीनोटाइप के कुछ नंबरों की वजह से पुनर्संयोजन की कम संख्या के परिणामस्वरूप मैप किए गए क्यूटीएल के कम रिज़ॉल्यूशन का परिणाम होता है। पारंपरिक प्रजनन के माध्यम से, आरआईएल के पास या समद्विबाहु लाइनों (एनआईएल) की शुद्ध लाइनों (समरूप रेखाओं) को बनाने के लिए अंतःसंक्रमण या सेल्फिंग की छह से आठ पीढ़ियों की आवश्यकता होती है, जबकि दो पीढ़ियों के लिए पुनर्संयोजन दर की घटनाओं की कम संभावना के साथ दो पीढ़ियों का निर्माण होता है। आरआईएल की आबादी। यह इस तथ्य के कारण हो सकता है कि डीएच लाइन केवल पुनर्संयोजन के एक दौर से गुजरती है, जबकि दूसरी ओर, आरआईएल पुनर्संयोजन के कई दौर से गुजरती है। एक एकल बीज वंश विधि का उपयोग करके** F2 **से समयुग्मक लाइनों का भी उत्पादन किया जा सकता है, जहां प्रत्येक** F2 **लाइन से एक बीज काटा जाता है और फिर** F8 **पीढ़ियों तक** F8 **से** F10 **पीढ़ी तक उच्च स्तर के साथ होमोजीगोसिटी के साथ लगभग सभी लोकी में उगाया जाता है। अंत में, एक परिवार-आधारित मानचित्रण आबादी के सदस्यों में लोकी के बीच पुनर्संयोजन की विभिन्न मात्राएं होंगी।**

इन सीमाओं से बचने के लिए, मैपिंग आबादी के भीतर मैपिंग रिज़ॉल्यूशन में सुधार को बहुपक्षीय आरआईएल का उपयोग करके इंटरक्रॉस की संख्या में वृद्धि करके नाटकीय रूप से सुधार किया जा सकता है। इस दृष्टिकोण का उपयोग करने में कई सकारात्मक विशेषताएं हैं। क्यूटीएल की मैपिंग के लिए एक उच्च पुनर्संयोजन दर (आरआईएल लाइनों) के मामले में उच्च घनत्व वाले मार्करों की आवश्यकता है और कसकर जुड़े मार्करों की पहचान करना है। यह स्थानीय स्तर पर विविधता को समझने के लिए भी मजबूत है। उन्नत आणविक प्रौद्योगिकियां जीबीएस ने उच्च एलील रिचनेस के साथ तीव्र और लागत प्रभावी जीनोटाइपिंग (सैकड़ों से हजारों मार्करों) की अनुमति दी जो जटिल लक्षणों के लक्ष्य क्षेत्र की पहचान करने के लिए क्यूटीएल मैपिंग को मजबूत और उपयोगी बनाते हैं।

जीनोम-वाइड एसोसिएशन स्टडी (**GWAS)**

GWAS **का उपयोग करके एसोसिएशन विश्लेषण जीनोम-फेनोटाइप एसोसिएशन और प्रेरक लोकी / जीन की पहचान के लिए प्रभावी और कुशलतापूर्वक उपयोग किया जा रहा एक शक्तिशाली उपकरण है। जीडब्ल्यूएएस में मूल परिदृश्य प्रत्येक मार्कर और ब्याज के एक फेनोटाइप के बीच संबंध की गणना करने के लिए है जो कि एक विविध संग्रह के असंबंधित लाइनों / व्यक्तियों (असंबंधित व्यक्तियों का अर्थ है कि दूर से संबंधित और विषम व्यक्तियों) के बीच स्कोर**

किया गया है। फसलों में जटिल लक्षणों के विच्छेदन में GWAS की तीव्रता और प्रभावशीलता का प्रदर्शन किया गया था और वर्तमान में उपलब्ध बड़ी आबादी और उच्च विवादास्पद अनुक्रमण तकनीक की मदद से मात्रात्मक लक्षणों के लिए प्रेरक लोकी / जीन (एस) की पहचान करने के लिए और अधिक कुशल बनने की उम्मीद की गई थी। उच्च-रिज़ॉल्यूशन मैपिंग को ऐतिहासिक पुनर्संयोजन की घटनाओं और GWAS में शामिल होने वाले अधिक से अधिक एलील संख्या के लिए भी जिम्मेदार ठहराया जा सकता है। एसोसिएशन मैपिंग पॉपुलेशन में, ऐतिहासिक संबंध जो ऐतिहासिक संबंध डेसीक्विलिब्रियम (एलडी, दर्जनों / सैकड़ों पीढ़ियों से अधिक) के साथ पीढ़ियों तक जमा हुए हैं, प्रतिनिधि अभिगम के बीच बने रहते हैं और एलडी के तेजी से क्षय के माध्यम से एसोसिएशन विश्लेषण के लिए संकल्प में सुधार हुआ है। एसोसिएशन मैपिंग आबादी के विपरीत, परिवार-आधारित आबादी, विशेष रूप से डीएच आबादी, एक सीमित संख्या में पुनर्संयोजन की घटना होने से अक्सर आबादी अपेक्षाकृत उत्पन्न होगी

GWAS की शक्ति को प्रभावित करने वाले महत्वपूर्ण कारक

वास्तविक एसोसिएशन का पता लगाने के लिए जीडब्ल्यूएएस की शक्ति कई कारकों द्वारा निर्धारित की जाती है, जिसे ध्यान में रखा जाना चाहिए जब आनुवंशिकीविद् और प्रजनक लक्ष्य लक्षणों के लिए जीडब्ल्यूएएस करते हैं, जीडब्ल्यूएएस के फायदे और सीमाएं जो निम्नानुसार वर्णित हैं:

पहला: फेनोटाइपिक भिन्नता।

कच्चे फेनोटाइपिक डेटा को आउटलेर्स से फ़िल्टर किया जाना चाहिए जो आगे के विश्लेषण के लिए शोर डेटा बिंदु हैं। इन बिंदुओं को रखने से फेनोटाइपिक डेटा को एक सामान्य वितरण से स्थानांतरित किया जा सकता है जिसे GWAS की सीमा माना जाता है जो बाद में प्राकृतिक विविधता विश्लेषण को प्रभावित कर सकता है। यह जानने का सरल तरीका है कि फेनोटाइपिक डेटा में कितने आउटलेयर हैं और वे प्रभावी हैं या नहीं, एक बॉक्सप्लॉट का उपयोग करना है जो आसानी से डेटा की कल्पना कर सकता है और चरम आउटलेयर को बाहर रखा जाना चाहिए। इस बीच, फेनोटाइपिक भिन्नता संघ के विश्लेषण का एक महत्वपूर्ण हिस्सा है और आउटलेयर को हटाने से इसे सार्थक तरीके से प्रभावित नहीं करना चाहिए। इसके अलावा, केवल मध्यम से उच्च आनुवांशिकता अनुमानों (निस्पंदन के बाद फेनोटाइपिक डेटा के लिए) के साथ लक्षण को GWAS में माना जाना चाहिए, क्योंकि आनुवांशिकता ने फेनोटाइप में कितना आनुवंशिक परिवर्तन और फेनोटाइप जीनोटाइप से जुड़ा है, इसका एक अच्छा संकेत है। कम व्यापक-अर्थ की आनुवांशिकता एक सीमित कारक है जिसने एसोसिएशन का पता लगाने के लिए GWAS की शक्ति को कम कर दिया है। स्थानों या वर्षों में दोहराए गए जीनोटाइप में एक मजबूत जीनोटाइप पर्यावरण इंटरैक्शन हो सकता है सर्वश्रेष्ठ रैखिक निष्पक्ष भविष्यवक्ता (BLUP) और सर्वश्रेष्ठ रैखिक निष्पक्ष अनुमानक (BLUE) जैसी कई विधियाँ हैं जिनका उपयोग जीनोटाइप पर्यावरण इंटरैक्शन पर विचार करने वाले फेनोटाइपिक के बेहतर अनुमान प्रदान करने के लिए स्थानों या वर्षों में किए गए फ़ेनोटाइपिक डेटा को समायोजित करने के लिए किया जा सकता है। असंबद्ध व्यक्तियों में संबंधित एसएनपी और फेनोटाइपिक लक्षणों के बीच संबंध को एसएनपी के विचरण के अनुमान से समझाया जाता है, जब एक जीडब्ल्यूएएस में उपयोग किया जाता है, जिसे तथाकथित एसएनपी-आधारित हेरिटैबिलिटी के रूप में भी जाना जाता है। इस तरह के विश्लेषण से आनुवंशिक भिन्नता का पता लगाने और जटिल लक्षणों के लिए आनुवंशिक संरचना को समझने में मदद मिलती है, इसके अलावा सबसे महत्वपूर्ण एसएनपी की पहचान करना जो भविष्य के प्रजनन कार्यक्रमों में शामिल किया जा सकता है।

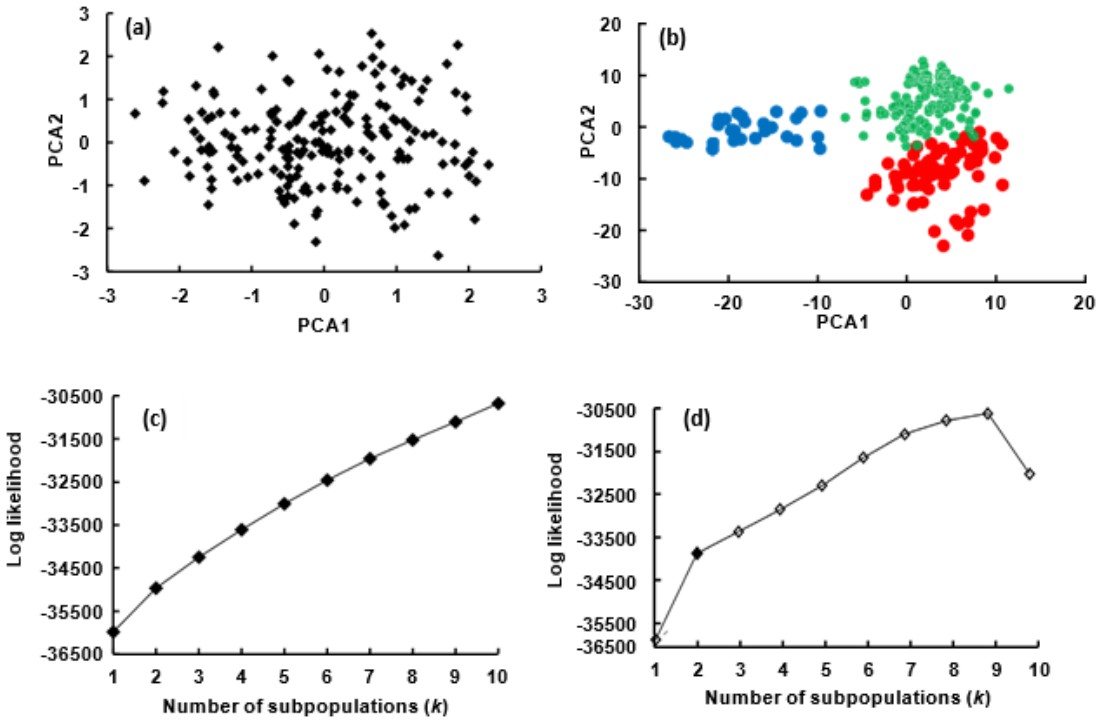
दूसरा: व्यक्तियों की संख्या।

सार्थक परिणाम प्राप्त करने के लिए जनसंख्या का आकार बहुत महत्वपूर्ण है। फेनोटाइपिक और जीनोटाइपिक भिन्नता के कुछ हिस्सों को परिभाषित करने के लिए जनसंख्या का आकार महत्वपूर्ण है; इसलिए जनसंख्या का आकार बढ़ने से बड़े प्रभाव के साथ सार्थक संघों की शक्ति में सुधार होगा, आबादी के भीतर एक स्वीकार्य आवृत्ति और दुर्लभ-रूपांतरों पर काबू पाया जा सकेगा। इस प्रकार, व्यक्तियों की कम संख्या एक नुकसान है जो GWAS की शक्ति को कम करता है। 100-500 व्यक्तियों की एक श्रृंखला की जरूरत है और प्रति के लिए उपयुक्त है भौगोलिक क्षेत्रों, जैविक स्थिति, विकास की आदत या शोधकर्ताओं द्वारा रुचि रखने वाले किसी भी लक्षण सहित आनुवंशिक पृष्ठभूमि पर विचार करने के लिए आबादी के व्यक्तियों को उनके अपेक्षित फेनोटाइपिक और जीनोटाइपिक भिन्नता के आधार पर चुना जा सकता है। चयनित व्यक्तियों को सांख्यिकीय विश्लेषण के माध्यम से उनकी विविधता की पुष्टि करने के लिए दोहराया जाना चाहिए। क्लस्टरिंग विश्लेषण सहित और एक सामान्य वितरण सुनिश्चित करने के लिए। अंत में, आगे के अनुसंधान उद्देश्यों के लिए पर्याप्त बीज की आवश्यकता होती है। जनसंख्या को गुणन और शुद्धता के लिए कई सेल्फिंग पीढ़ियों के लिए प्राथमिकता के साथ कम से कम एक बढ़ते मौसम के लिए उगाया और अलग किया जाना चाहिए। एकल बीज वंश। जनसंख्या व्यक्तियों के सावधानीपूर्वक चयन से बड़े आनुवंशिक परिवर्तन हो सकते हैं और

सच्चे उपन्यास संघ संकेतों का पता लगा सकते हैं जिनका उपयोग आगे प्रजनन और आनुवंशिकी पहलुओं के लिए किया जा सकता है।

तीसरा: जनसंख्या संरचना।

यह एक सांख्यिकीय दृष्टिकोण है जिसका उद्देश्य विश्लेषण और परिणाम व्याख्या के दौरान सावधानीपूर्वक विचार किए जाने वाले प्रवेश और ऐतिहासिक संरचना के कारण आबादी के भीतर व्यक्तियों के बीच संबंधित सहसंबंध की गणना करना है। शोधकर्ताओं द्वारा एसोसिएशन विश्लेषण के लिए आबादी का चयन भौगोलिक या विकास की आदत, आदि के आधार पर संरचना उत्पन्न करता है। यह एक विशिष्ट आनुवंशिक भिन्नता और संघ विश्लेषण के अंतिम उपयोग पर प्रभाव डालता है। यह GWAS **विश्लेषण में प्रमुख सीमा है क्योंकि सभी व्यक्ति आनुवंशिक स्तर पर समान रूप से एक दूसरे से संबंधित नहीं हैं। जनसंख्या संरचना के सुधार को नजरअंदाज करने से जीनोटाइप और अभिरुचि के बीच सहज जुड़ाव होता है। संरचना कार्यक्रम जनसंख्या संरचना को परिभाषित करने के लिए एक कम्प्यूटेशनल रूप से गहन विधि है और फिर जनसंख्या के भीतर समूहों के अनुपात (उप संख्याओं की अज्ञात संख्या) को तथाकथित क्यू मैट्रिक्स कहा जाता है और फिर अनुमान लगाया जाता है कि कौन सा व्यक्ति किस उप-समूह से संबंधित है। सॉफ्टवेयर जनसंख्या संरचना की व्याख्या करने के लिए जीनोटाइप से मल्टीकोकस डेटा का उपयोग करके अत्यधिक सटीक क्लस्टरिंग का उत्पादन करता है। समूहों की संख्या को परिभाषित करने और समूहों में व्यक्तियों को कैसे निर्दिष्ट किया जाए, इसकी वजह से संरचित संघों को हटाना हमेशा जनसंख्या संरचना को नियंत्रित करने के लिए पर्याप्त नहीं है। इसके अलावा, संरचना विश्लेषण समय लेने वाली गहन कम्प्यूटेशनल विश्लेषण की आवश्यकता हो सकती है। वैकल्पिक रूप से, प्रिंसिपल कंपोनेंट एनालिसिस (**PCA) **का उपयोग करने वाला** EIGENSTRAT **विधि प्राइस एट अल द्वारा विकसित एक और सांख्यिकीय दृष्टिकोण है। संरचना को नियंत्रित करने के लिए आयामी जीनोटाइप डेटा को कम करने के लिए जनसंख्या की संरचना को गिना जाता है। विधि आनुवंशिक भिन्नता को कम करने के लिए जीनोटाइपिक डेटा पर विचार करती है जिसे बहुत कम आयामों द्वारा समझाया जा सकता है। यू एट अल। रिश्तेदारी मैट्रिक्स** (K) **नामक जोड़ीदार संबंधित मैट्रिक्स के माध्यम से संबंधित कई स्तरों के लेखांकन के माध्यम से संयमी संघों को नियंत्रित करने के लिए एक मिश्रित-मॉडल दृष्टिकोण विकसित किया।** K **जीनोटाइपिक जानकारी का उपयोग करने वाले व्यक्तियों के जोड़े के बीच संबंधितता की गणना कर सकता है। व्यक्तियों के बीच संबंधों का उच्च मूल्य उच्च आनुवंशिक समानता को इंगित करता है उदा। एक ही भौगोलिक क्षेत्र के व्यक्तियों के बीच की प्रवृत्ति, जिन्हें समूह में जोड़ा जा सकता है। अधिकांश अध्ययन अपने परिणामों की पुष्टि करने के लिए दोनों तरीकों (**STRUCTURE **और** PCA) **का उपयोग करते हैं। प्रिंसिपल कंपोनेंट विश्लेषण को** PCA1 **और** PCA2 **के स्कैटर प्लॉट में प्रस्तुत किया गया है, जो कि उनके जीनोटाइपिक डेटा के आधार पर व्यक्तियों के बीच कुल भिन्नता का सबसे अधिक उपयोग करता है। यदि जीनोटाइप को प्लॉट में बेतरतीब ढंग से वितरित किया जाता है और कोई स्पष्ट समूह नहीं बनता है, तो आबादी में कोई जनसंख्या संरचना नहीं होती है, और इसके विपरीत (छवि** 1 **ए और बी)। संरचना सॉफ्टवेयर में, पॉपुला टियन संरचना डेल्टा** k **के खिलाफ प्रस्तावित उप-योगों की साजिश रचकर निर्धारित की जाती है। हालाँकि, यदि उपसमूहों की संख्या को दो उप-योगों में असाइन किया गया है, तो जनसंख्या में दो संभावित उप-योग या कोई जनसंख्या संरचना हो सकती है क्योंकि** STRUCTURE **पहले उप-युग्मन के लिए डेल्टा** k **का अनुमान नहीं लगाता है। जनसंख्या संरचना की उपस्थिति या अनुपस्थिति को किसी अन्य भूखंड द्वारा निर्धारित किया जा सकता है जिसमें लॉग के खिलाफ कई उप-संरचनाएँ प्लॉट की जाती हैं- कोई जनसंख्या संरचना (छवि** 1 **सी) के मामले में, उप-योगों की संख्या में वृद्धि के साथ लॉग-लाइबिलिटी लगातार बढ़ जाती है। यदि दूसरी ओर, लॉग-लाइबिलिटी,** k = 2 **(छवि** 1d**) के बाद लगातार बढ़ जाती है, तो इस आबादी को दो संभावित उप-वर्गों में विभाजित किया जा सकता है।** STRUCTURE **हारवेस्टर** (http://taylor0.biology.ucla.edu/structureHarvester/) **एक बहुत ही उपयोगी वेबसाइट है जिसमें** STRUCTURE **के आउटपुट परिणाम को संकुचित और अपलोड किया जा सकता है। सॉफ्टवेयर जनसंख्या पर जानकारी प्रदान करता है और तालिका और आंकड़ों में प्रस्तावित आबादी के लिए सबसे अच्छा कश्मीर है।**



चित्रा। 1. **जनसंख्या संरचना का दृश्य और आबादी के भीतर उप-जनसंख्या की संख्या। कोई स्पष्ट जनसंख्या संरचना (ए), जबकि जनसंख्या अच्छी तरह से संरचित थी (बी)।** STRUCTURE **रन से** k (**फ़ंक्शन की संख्या / क्लस्टर / उप-जनसंख्या) के रूप में संभाव्यता डेटा लॉग करें। उप-योगों की संख्या नहीं** (c), **जबकि दो उप-योगों को** (d) **में दिखाया गया है। प्रत्येक रंग (ए और बी) एक उपसमूह का प्रतिनिधित्व करता है और प्रत्येक डॉट एक परिग्रहण / व्यक्ति का प्रतिनिधित्व करता है। पीसीए, प्रमुख घटक विश्लेषण।**

चौथा: एलील आवृत्ति।

एक बहुत ही महत्वपूर्ण कारक, जो GWAS **की शक्ति को प्रभावित करता है, यदि एलील आबादी में कुछ व्यक्तियों में मौजूद हैं। दुर्लभ एलील संकल्प शक्ति की कमी की ओर जाता है। इसलिए, एसोसिएशन का पता लगाने पर आवृत्ति वितरण और विश्लेषण प्रभाव को नियंत्रित करें। कार्यात्मक आवृत्तियों का पता लगाना मुश्किल है जो कम आवृत्ति पर मौजूद होते हैं जब तक कि उनके फेनोटाइप पर उच्च प्रभाव न हों। एलील आवृत्ति की अनदेखी** GWAS **आउटपुट को भ्रमित कर सकती है। जीडब्ल्यूएएस के अधिकांश अध्ययन केवल आम वेरिएंट पर केंद्रित हैं और** 5% **पर प्रमुख एलील आवृत्ति है। इस दृष्टिकोण का अर्थ है कि** 200 **व्यक्तियों की आबादी में,** 10 **व्यक्तियों या उससे कम में मौजूद एलील का पता नहीं लगाया जाएगा क्योंकि यह** 5% **से कम पर मामूली एलील आवृत्ति** (MAF) **के साथ एक दुर्लभ संस्करण है। दुर्भाग्य से, दुर्लभ एलील व्यक्तियों के एक विशिष्ट समूह में प्राकृतिक भिन्नता को समझा सकते हैं जो जैविक अध्ययन के अलावा आगे प्रजनन और आनुवंशिकी के लिए महत्वपूर्ण है।**

पांचवां: एलडी

LD **एक और बिंदु है जिसे विश्लेषण के दौरान विचार किया जाना है, विशेष रूप से उच्च जुड़े** SNPs **के अंतराल को परिभाषित करने के लिए जो सबसे महत्वपूर्ण लोकी को परिभाषित करने का कारण बन सकता है। अलग-अलग लोकी में एलील्स के बीच गैर-यादृच्छिक जुड़ाव को अनदेखा करने का अर्थ है कि कारण और गैर-कारण वाले एलील दोनों को आगे के विश्लेषणों में शामिल किया जाएगा, जिसमें संभवत: गलत संघों के लिए ड्राइविंग की संभावना है।** एलडी लोकी के बीच की दूरी का पता लगाने के लिए एक संकेतक है, जो पूरे जीनोम स्कैन के लिए आवश्यक मार्करों की संख्या को खोजने के लिए महत्वपूर्ण है, अर्थात उच्च एलडी मान का मतलब है जीनोम को कवर करने के लिए मार्करों की कम संख्या की आवश्यकता होती है। एक लंबी दूरी के एलडी में झूठी एसोसिएशन की संभावना बढ़ जाती है और इसलिए, एसोसिएशन विश्लेषण की शुरुआत में एलडी की गणना आवश्यक है। एलडी के गुणांक का उपयोग दो लोकी से जुड़े होने और उत्परिवर्तन और पुनर्संयोजन के इतिहास को साझा करने की संभावना के मूल्य को मापने के लिए किया जाता है। इस विश्लेषण में हमेशा एक असमान मैट्रिक्स शामिल होता है जो LD यानि r2 और D '[10] को

मापने के लिए सबसे सामान्य दो आँकड़ों का उपयोग करके लोकी के बीच युग्मक गणनाओं को दिखाता है। पौधों में कई एलडी विश्लेषणों के अनुसार, आर 2 एक मजबूत मूल्य है यह अनुमान लगाने के लिए कि लोकी ब्याज की क्यूटीएल के साथ कैसे संबंधित है, जबकि डी 'छोटे जनसंख्या आकार और कम एलील आवृत्तियों से अधिक प्रभावित होता है। क्योंकि एलडी का उपयोग लोकी (आर 2 या डी ',> 0) के बीच संघ मूल्य की गणना करने के लिए किया जाता है। यह कारणात्मक एसएनपी के साथ फेनोटाइपिक भिन्नता को जोड़ने के लिए महत्वपूर्ण है। एसएनपी के बीच एलडी, प्रेरक स्थान सहित (एलडी के भीतर) एक सांख्यिकीय विश्लेषण में विचार किया जाना चाहिए, जो यह दिखा सकता है कि एलडी के भीतर प्रत्येक एसएनपी फेनोटाइपिक भिन्नता के साथ महत्वपूर्ण रूप से जुड़ा हुआ है या नहीं। यहाँ, हम सभी एसएनपी को थ्रेशोल्ड (कुछ मामलों में सभी एसएनपी) से ऊपर विचार करने के लिए इस तरह के विश्लेषण में प्रस्तावित करते हैं कि कौन अधिक प्राकृतिक फेनोटाइपिक भिन्नता की व्याख्या कर सकता है क्योंकि यह ज्ञात है कि सभी एसएनपी अत्यधिक संबद्ध एसपीपी पर अत्यधिक महत्वपूर्ण प्रभाव नहीं डालते हैं फेनोटाइप। एसएनपी जो आरडी> 0.2 के साथ एलडी में हैं, उन्हें सांख्यिकीय विश्लेषण में माना जाना चाहिए जो विशेष रूप से क्यूटीएल के लिए करणीय लोकी का पता लगाने के लिए उपयोगी हो सकता है जो सेंट्रोमीटर क्षेत्र में स्थित हैं। एसएनपी के प्रत्येक जोड़े के बीच एलडी के अनुमान के रूप में गणना की गई आर 2 की एक और विशेषता यह है कि यह महत्वपूर्ण जानकारी देता है यदि महत्वपूर्ण एसएनपी का एक समूह एक साथ या एक ही क्यूटीएल या व्यक्तिगत क्यूटीएल का प्रतिनिधित्व करता है। एक ही गुणसूत्र पर स्थित महत्वपूर्ण एसएनपी को देखकर, यदि दो एसएनपी के बीच r2 मान अधिक है, तो ये एसएनपी संभवतः एक ही क्यूटीएल का प्रतिनिधित्व करते हैं और एक साथ विरासत में मिलते हैं, जबकि यदि मूल्य कम है, तो दो महत्वपूर्ण एसएनपी शायद दो अलग-अलग क्यूटीएल का प्रतिनिधित्व करते हैं।

GWAS कैसे काम करता है

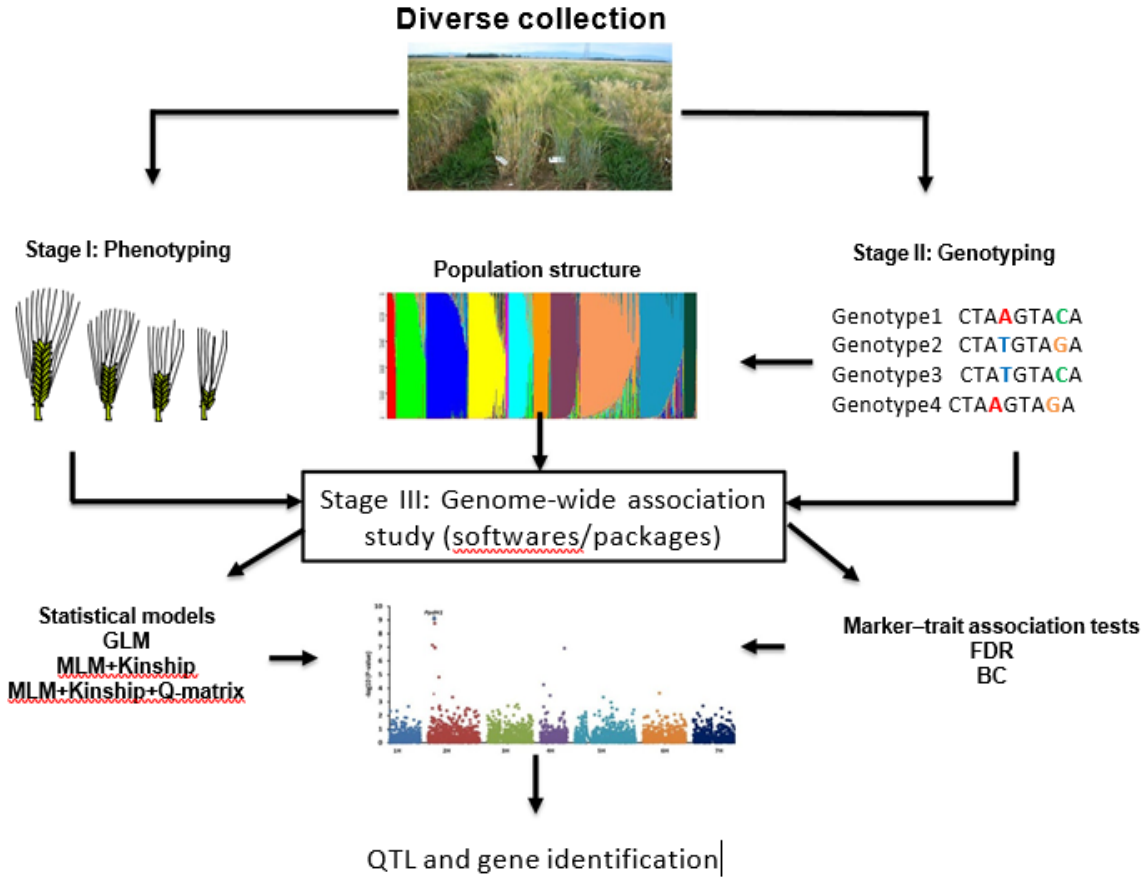
जीडब्ल्यूएएस प्रयोग करने के लिए, पहला कदम जनसंख्या के आकार (न्यूनतम 100 व्यक्तियों) का पूर्ण विचार के साथ अध्ययन की आबादी का चयन करना है, जहां बीवियों के प्रभावों से बचने के लिए यथासंभव अधिक से अधिक व्यक्तियों की संख्या बढ़ाना है। जब व्यक्तियों की संख्या छोटी होती है, तो फेनोटाइपिक विचरण की अधिकता १०० [११]। फिर, एक सफल GWAS प्रयोग करने के लिए तीन महत्वपूर्ण चरण हैं (चित्र 2);

स्टेज I वह फेनोटाइपिंग है जिसमें अध्ययन के उद्देश्यों के आधार पर सभी जीनोटाइप किसी विशेष गुण या लक्षणों के समूह के लिए फेनोटाइप किए जाने चाहिए। जीनोटाइप-फेनोटाइप एसोसिएशनों का पता लगाने के लिए सटीक फेनोटाइपिंग एक बहुत ही महत्वपूर्ण बिंदु है। फेनोटाइपिंग को प्रतिकृति और / या स्थानों और / या वर्षों में दोहराया जाना चाहिए। कच्चे डेटा के लिए व्यापक-अर्थ हेरिटैबिलिटी की गणना की जानी चाहिए (ध्यान दें, इन सभी कारकों सहित जीई इंटरैक्शन पर विचार करते हुए आउटलेयर को हटाने के बाद गणना की जानी चाहिए)। उच्च आनुवांशिकता एक संकेतक है कि विशेषता ज्यादातर आनुवंशिक रूप से नियंत्रित होती है जो एसोसिएशन संकेतों का पता लगाने के लिए महत्वपूर्ण है। फिर, फेनोटाइपिक डेटा का उपयोग माध्य यानी BLUE या BLUP का अनुमान लगाने के लिए किया जा सकता है। क्योंकि पौधों में फेनोटाइपिक डेटा अत्यधिक असंतुलित होते हैं, जीनोटाइपिक मूल्यों का अनुमान ज्यादातर मिश्रित प्रभावों का उपयोग करके निश्चित प्रभावों (यानी BLUE) के रूप में गणना की जाती है।

स्टेज II (चित्र। 2)

जीनोटाइपिंग है जिसमें डीएनए आणविक मार्करों का उपयोग करके जीनोटाइपिंग के लिए फ़िनोटाइप किए गए व्यक्तियों का एक ही सेट इस्तेमाल किया जाना चाहिए। जीबीएस जीनोटाइपिंग में उपयोग की जाने वाली सबसे लगातार विधि है क्योंकि यह कई एसएनपी मार्करों को सस्ते में उत्पन्न करता है जो फसल जीनोम (जैसे गेहूं, जौ, आदि) को कवर करते हैं। जीबीएस द्वारा उत्पन्न एसएनपी को लापता डेटा, हेटेरोज़ायोसिटी और मामूली एलील आवृत्ति के आधार पर फ़िल्टर किया जाना चाहिए। GWAS चलाने से पहले, बेहतर GWAS मॉडल का चयन करने के लिए जनसंख्या संरचना का परीक्षण किया जाना चाहिए। सामान्य रैखिक मॉडल (GLM) और मिश्रित रैखिक मॉडल (MLM) सांख्यिकीय मॉडल हैं जिन्हें अक्सर GWAS (चित्र 2) प्रदर्शन के लिए प्रस्तावित किया जाता है। जीएलएम जनसंख्या संरचना को ध्यान में नहीं रखता है। इसलिए, जीएलएम का उपयोग आबादी में किया गया था जिसमें फैबा बीन, विकिया फैबा एल और चावल में जनसंख्या संरचना नहीं थी। दूसरी ओर, एमएलएम, अपने मॉडल (रिश्तेदारी या रिश्तेदारी क्यू मैट्रिक्स पीसीए) में जनसंख्या संरचना पर विचार करता है। अंत में, फेनोटाइपिक और जीनोटाइपिक डेटा को उपयुक्त सॉफ्टवेयर (जैसे TASSEL) का उपयोग करके संयोजित किया जाता है, जिसके द्वारा GWAS मॉडल चुने जाने के बाद किसी विशेष विशेषता से जुड़े एलील्स का पता लगाया जा सकता है (स्टेज III: चित्रा 2)। विशेष रूप से

बिना किसी पूर्व सूचना के उन लोगों के लिए जीनोटाइपिंग से पहले फेनोटाइपिंग का आयोजन किया जाना अत्यधिक अनुशंसित है। उदाहरण के लिए, यदि जनसंख्या में 400 जीनोटाइप शामिल हैं जो विभिन्न क्षेत्रों से एकत्र किए गए थे और उद्देश्य उन्हें एक विशेष वातावरण में परीक्षण करना है। यह संभव है कि फेनोटाइपिंग वातावरण के खराब अनुकूलन के कारण कई जीनोटाइप खो सकते हैं। इसलिए, समय और धन (जीनोटाइपिंग के लिए) पहले उस जनसंख्या की फेनोटाइपिक विविधता का परीक्षण करके बचाया जा सकता है। संबद्ध मार्कर-विशेषता का महत्व (दहलीज जैसे-log10 p-value> 3) आमतौर पर झूठी खोज दर (FDR) या बोनफेरोनि सुधार (BC) द्वारा निर्धारित किया जाता है जिसे कई तुलनाओं के रूप में परिभाषित किया जा सकता है जो परीक्षण करने के लिए फिट हो सकते हैं GWAS में मार्करों के लाखों लोगों के सैकड़ों का महत्व। बीसी के लिए, महत्वपूर्ण स्तर को प्रत्येक स्थान पर परीक्षणों (मार्कर) की संख्या से विभाजित किया गया है। बीसी विधि का गहन रूप से उपयोग किया जाता है जैसे एक बार में कई लक्षणों के लिए महत्वपूर्ण मार्करों की सीमा को परिभाषित करना। नतीजतन, एक निश्चित बीसी पी-मूल्य। झूठी खोज दर (एफडीआर) एक और परीक्षा है जो महत्वपूर्ण कहे जाने वालों के बीच वास्तविक वास्तविक परिणामों की संख्या का अनुमान प्रदान करती है। इस परीक्षण में, GWAS से उत्पन्न सभी मार्करों के पी-मूल्यों को आरोही क्रम में क्रमबद्ध किया गया है। फिर, प्रत्येक स्थान पर प्रत्येक पी-मूल्य को एक रैंक (आर - जैसे 1, 2, 3, ... 100,000) दिया जाता है। एफडीआर की गणना प्रत्येक गुण के लिए स्वतंत्र रूप से की जाती है, जो फसल के पौधों में विकासात्मक और एग्रोनोमिक लक्षणों के आनुवंशिक कारकों का अध्ययन करने में अधिक शक्तिशाली बनाता है। सभी लक्षणों के लिए निर्धारित पी-मूल्य (बीसी) की तुलना में, मार्करों और लक्षणों के आधार पर, पी-मूल्य (एफडीआर) अधिक लचीला और परिवर्तित होता है। इसलिए, एफडीआर बीसी की तुलना में कम रूढ़िवादी है और स्वतंत्र रूप से प्रत्येक निशान के लिए अत्यधिक जुड़े मार्करों का पता लगाने के लिए फसल संयंत्र संघ के अध्ययन में उपयोग करने की सिफारिश की जाती है। दोनों परीक्षणों में और प्रत्येक स्थान पर, यदि एफडीआर या बीसी का पी-मान पीडब्ल्यू मूल्य से कम या बराबर है, तो मार्कर के जीडब्ल्यूएएस से उत्पन्न, एसोसिएशन सही है और मार्कर विशेषता के साथ जुड़ा हुआ है। मार्कर-विशेषता संघ का परीक्षण 0.01 और 0.05 [12] के महत्व स्तर पर किया जा सकता है। हालांकि, कुछ एसोसिएशन एनालिसिस स्टडीज ने एफडीआर का 20% महत्व स्तर पर उपयोग करके मार्कर-ट्रिट एसोसिएशन का परीक्षण किया क्योंकि यह मामूली प्रभावों के साथ महत्वपूर्ण मार्करों का पता लगा सकता है। GWAS में मार्कर-विशेषता संघों के लिए महत्व स्तर का निर्धारण अध्ययन पर आधारित है, जो महत्वपूर्ण आनुवंशिक और आणविक अध्ययनों के लिए उम्मीदवार लोकी / जीन की पहचान करने के लिए एक विशेषता या कम FDR के आनुवंशिक वास्तुकला की पूरी तस्वीर की जांच करने के लिए उच्च FDR का उपयोग कर सकता है।



चित्रा। 2. एक सफल GWAS प्रयोग करने के लिए सबसे महत्वपूर्ण तीन चरण। स्टेज I: फेनोटाइपिंग, स्टेज II: जीनोटाइपिंग और

स्टेज III: जीनोम-वाइड एसोसिएशन स्टडीज जिसमें सांख्यिकीय मॉडल, कई-परीक्षण विश्लेषण और क्यूटीएल और जीन पहचान के लिए सॉफ्टवेयर / पैकेज शामिल हैं।

GWAS (TASSEL, PLINK, और R (GAPIT)) के प्रदर्शन के लिए सॉफ्टवेयर

GWAS कई सॉफ्टवेयर सांख्यिकीय पैकेज (स्टेज III: अंजीर 2) का उपयोग करके किया जा सकता है। यहां, हम सबसे महत्वपूर्ण एसोसिएशन विश्लेषण सॉफ़्टवेयर पैकेजों पर ध्यान केंद्रित करते हैं जो अक्सर उपयोग किए जाते हैं। TASSEL (एसोसिएशन, इवोल्यूशन और लिंकेज द्वारा विशेषता विश्लेषण) पौधों में GWAS के लिए सबसे आम सॉफ्टवेयर है। इसमें GLW और MLM [13] सहित GWAS के प्रदर्शन के लिए कई शक्तिशाली सांख्यिकीय तरीके शामिल हैं। TASSEL रिश्तेदारी और पीसीए का उपयोग करके जनसंख्या संरचना का विश्लेषण कर सकता है। LD को TASSEL में भी शामिल किया गया है। TASSEL (TASSEL 5.0) का नया संस्करण आनुवंशिक विविधता का विश्लेषण कर सकता है और GBS डेटा से SNP कॉलिंग कर सकता है। दिलचस्प बात यह है कि सॉफ्टवेयर में कई विज़ुअलाइज़िंग टूल शामिल हैं, जिनका उपयोग डेटा को पेश करने के लिए किया जा सकता है, जैसे कि पीसीए, एलडी, मैनहट्टन प्लॉट ऑफ़ जीडब्ल्यूएएस के परिणाम, आनुवांशिक दूरी के लिए हीट मैप, फेनोटाइपिक विचरण के अलावा पुरातनता का उपयोग करते हुए एक फ़ाइलेगनेटिक ट्री के बारे में बताया गया है। मार्करों (R2) द्वारा। नए संस्करण में कुछ उपयोगी डेटा सारांश भी शामिल हैं, जो प्रत्येक गुणसूत्र पर जीनोटाइप, मार्कर, विषमयुग्मजी, लापता डेटा और मार्करों की संख्या पर एक शोधकर्ता के लिए एक त्वरित दृश्य प्रदान करते हैं। TASSEL के पुराने संस्करण जैसे TASSEL v.2.1 किसी भी प्रकार के डीएनए मार्कर (जैसे एसएनपी, एसएसआर, एएफएलपी, आरएपीडी, आदि) को स्वीकार कर सकते हैं। TASSEL v.5.0 केवल SNP मार्करों को स्वीकार करता है। TASSEL मुफ्त सॉफ्टवेयर है और इसे http: //www.maizegenet- ics.net/tassel से डाउनलोड किया जा सकता है।

PLINK फेनोटाइप्स और जीनोटाइप्स [14] के एक बड़े डेटासेट के अध्ययन की अनुमति देता है। यह मुफ्त सॉफ्टवेयर है जिसे http://zzz.bwh.harvard.edu/plink/ से डाउनलोड किया जा सकता है। यह कई विशेषताओं और विशेषताओं को प्रदान करता है, जिनमें से, PLINK जनसंख्या स्तरीकरण का पता लगाने, बुनियादी संघ परीक्षण, मेटा-विश्लेषण, और कुछ अन्य परीक्षण जैसे कि एसोसिएशन के लिए जीन-आधारित परीक्षण और एपिस्टासिस के लिए स्क्रीनिंग के लिए विश्लेषण करता है। मैनहट्टन भूखंड, क्यू-क्यू साजिश, और बहुआयामी स्केलिंग (जनसंख्या संरचना के लिए) के लिए चित्रमय चित्रों को चित्रित किया जा सकता है। इसके अलावा, SNP मार्करों के बीच GWAS और LD के परिणाम PLINK द्वारा उत्पादित तालिकाओं में प्रस्तुत किए जा सकते हैं। R सांख्यिकीय पर्यावरण मुक्त सॉफ़्टवेयर (https://www.r-project.org/) में हालिया प्रगति GWAS प्रदर्शन के लिए कई उपयोगी पैकेज प्रदान करती है। जीनोम एसोसिएशन और भविष्यवाणी एकीकृत उपकरण (GAPIT) एक उपयोगी आर पैकेज है जो GWAS और जीनोमिक चयन करता है। जीएपीआईटी के मुख्य लाभ हैं: यह बड़ी मात्रा में डेटा (एसएनपी और जीनोटाइप) को संभाल सकता है और यह सांख्यिकीय शक्ति से समझौता किए बिना कम्प्यूटेशनल समय को कम करता है। पैकेज में कई सांख्यिकीय विधियाँ शामिल हैं जैसे MLM, पहले से निर्धारित जनसंख्या पैरामीटर (P3D), और कुशल मिश्रित-मॉडल एसोसिएशन (EMMA)। जीडब्ल्यूएएस के परिणामों को मैनहट्टन भूखंडों, क्वांटाइल-क्वांटाइल (क्यूक्यू) भूखंडों और एक तालिका द्वारा चित्रित किया जा सकता है, जिसमें पी-मूल्य, मामूली एलील आवृत्ति, नमूना आकार, फेनोटाइपिक विचरण, मार्करों द्वारा समझाया गया है R2 और समायोजित पी-मान एक झूठी खोज के बाद। मूल्यांकन करें। इसी तरह, परिजन जहाज के परिणाम एक हीट मैप और एक तालिका में प्रस्तुत किए जाते हैं। इसके अलावा, अलग-अलग संपीड़न स्तरों पर हेरिटैबिलिटी अनुमान और संभावना फ़ंक्शन को ग्राफ़ में उत्पादित किया जा सकता है। उपरोक्त सुविधाओं के कारण, GAPIT जौ या गेहूं जैसे अन्य अनाज में संघ विश्लेषण के लिए सबसे शक्तिशाली और उपयोगी उपकरण बन जाता है।

GWAS के आउटपुट परिणाम

प्रत्येक सॉफ्टवेयर प्रोग्राम GWAS के आउटपुट परिणाम के रूप में थोड़ा अलग पैरामीटर देता है। TASSEL सॉफ्टवेयर कई मापदंडों के उत्पादन का एक अच्छा उदाहरण है जो लक्ष्य विशेषता के आनुवंशिक आधार को विच्छेदित करने में मदद करता है। इन मापदंडों में प्रत्येक एसएनपी का पी-मूल्य शामिल है, जो विशेषता के साथ महत्व को निर्धारित करने के लिए महत्वपूर्ण है, आर 2 (फेनोटाइपिक भिन्नता मार्कर द्वारा समझाया गया) जो यह निर्धारित करता है कि क्या महत्वपूर्ण एसएनपी एक मामूली या प्रमुख क्यूटीएल है, और महत्वपूर्ण एसएनपी के एलील प्रभाव (बढ़ा या घटाया)। मेन आउटपुट को मैनहट्टन प्लॉट में प्रस्तुत किया जा सकता है, जो कि जीनोमिक पैमाने पर, GWAS में उपयोग किए जाने वाले सभी मार्करों के पी-मूल्यों को दर्शाता है। एक्स-अक्ष गुणसूत्र द्वारा जीनोमिक क्रम और गुणसूत्र पर स्थिति का प्रतिनिधित्व करता है, जबकि, वाई-अक्ष प्रत्येक मार्कर के पी-मूल्य के दशमलव -1010 (दशमलव बिंदु प्लस एक के बाद शून्य की संख्या के बराबर) का प्रतिनिधित्व करता है। संबंधित

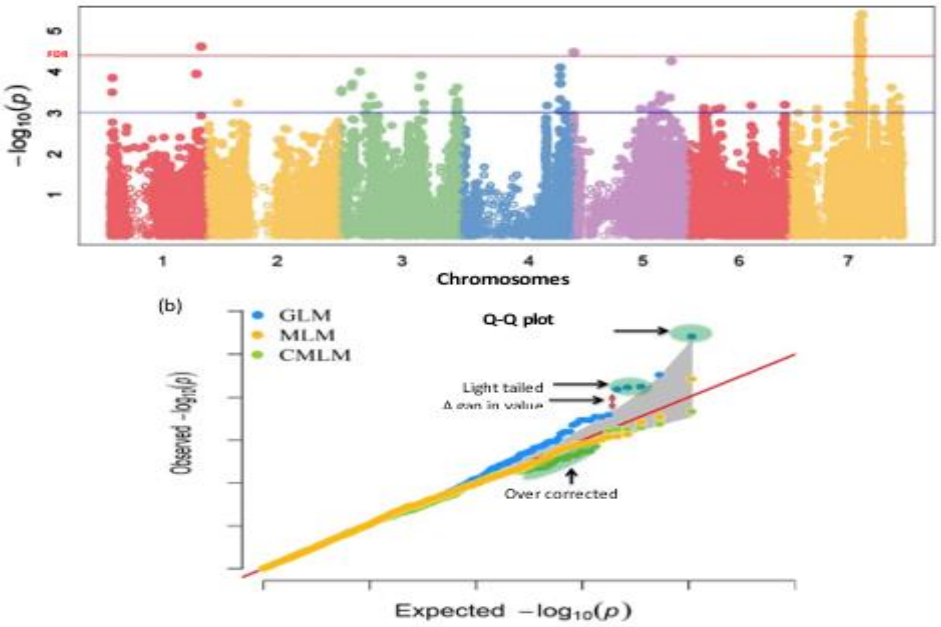
महत्वपूर्ण एसएनपी (सबसे महत्वपूर्ण पी-मान), क्यूटीएल का प्रतिनिधित्व करते हुए मैनहट्ट प्लॉट (चित्रा 3 ए) पर एक मजबूत संकेत के रूप में दिखाते हैं। -Log10 (p-value) की सीमा विश्वास मूल्य पर तय की जा सकती है, जिसका –log10 3 सबसे आम और विश्वसनीय मूल्य है (चित्र। 3A)। आगे के विश्लेषण के लिए, दहलीज को कई तुलनात्मक विश्लेषण का उपयोग करके पुनर्गणना की जा सकती है जो एसएनपी के पी-मूल्य को अधिक मजबूत और विश्वसनीय बनाता है (चित्रा 3 ए)। GWAS में एक अन्य महत्वपूर्ण ग्राफ QQ प्लॉट है जो देखे गए और अपेक्षित पी-मानों के बीच के संबंध को दर्शाता है। यह अशक्त परिकल्पना से प्रत्येक एसएनपी के देखे गए पी-मूल्य के विचलन को दर्शाता है। QQ प्लॉट का उपयोग GWAS स्टेटिक रूप से देखे गए मॉडल के बीच अपेक्षित मूल्यों की तुलना करने के लिए किया जा सकता है, यह दिखाने के लिए कि GWAS में मॉडल कितनी अच्छी तरह से जनसंख्या संरचना और पारिवारिक संबंधितता पर विचार करता है और फिर इसे लागू किया जा सकता है, उदाहरण के लिए, MLM GLM या CMLM मॉडल की तुलना में (चित्र 3 बी)। विकर्ण या मानक रेखा (छवि 3 बी में लाल) से पता चलता है कि क्या बिंदु पूरी तरह से मेल खाते हैं या विचलन वाले हैं जो वितरण को दर्शाते हैं। ग्रे क्षेत्र मूल्यों के लिए 95% विश्वास क्षेत्र दिखाता है। यह उम्मीद की जाती है कि क्यूक्यू भूखंड में अधिकांश डेटा बिंदु विकर्ण रेखा पर स्थित होंगे क्योंकि वे विशेषता से जुड़े नहीं हैं। जबकि इस पंक्ति के विचलन से यह संकेत मिलता है कि मॉडल जनसंख्या संरचना को पर्याप्त रूप से नियंत्रित नहीं करता है जिसे व्याख्यात्मक संघों के रूप में व्याख्या किया जा सकता है।

तीन मुख्य संभावित क्यूक्यू भूखंड हैं, जिनमें से प्रत्येक का अपना अर्थ है:

(1) देखे गए मान अपेक्षित मूल्यों के अनुरूप हैं, सभी बिंदु (देखे गए बनाम पी-मान) विकर्ण रेखा के पास या विश्वास अंतराल के भीतर, ग्रे हाइलाइट किए गए क्षेत्र (छवि 3 बी) के बहुत पास हैं।

(2) महत्वपूर्ण एसएनपी (देखा गया पी-मान अत्यधिक हैं और शून्य परिकल्पना के अनुसार p-मानों से भिन्न रूप से भिन्न हैं) y- अक्ष (चित्र 3B) की ओर बढ़ते हैं।

(3) यदि अंकों या अस्पष्ट प्रवृत्ति का प्रारंभिक पृथक्करण होता है, तो इसका अर्थ है कि परिणाम अनियंत्रित जनसंख्या संरचना या / और फेनोटाइपिक डेटा की खराब गुणवत्ता के कारण हो सकते हैं। इस मामले में, अधिकांश अत्यधिक विचलन वाले एसएनपी को एक झूठा संघ और अन्य विचारों के रूप में प्रस्तुत किया जाता है (जैसे कि जनसंख्या संरचना में सुधार, फेनोटाइपिक डेटा सुधार) (छवि 3 बी) की आवश्यकता होती है।



चित्रा। 3. GWAS के आउटपुट परिणाम। मैनहट्टन भूखंड (ए)। क्षैतिज-अक्ष जौ गुणसूत्रों पर मार्करों की स्थिति का प्रतिनिधित्व करता है और ऊर्ध्वाधर-अक्ष मार्कर-विशेषता संघ के -10 (पी-मान) का प्रतिनिधित्व करता है। प्रत्येक डॉट मार्कर को दर्शाता

है। क्षैतिज ब्लू-लाइन, -log10 (0.001) की दहलीज का प्रतिनिधित्व करती है और लाल-रेखा झूठी-खोज दर (FDR) से गुजरती -log10 (p- मान) की दहलीज का प्रतिनिधित्व करती है। विभिन्न GWAS मॉडल (b) के क्वांटाइल-क्वांटाइल (QQ) प्लॉट। प्लॉट प्रत्येक मार्कर (वोट) के अपेक्षित बनाम देखे गए -log10 (पी-मूल्य) को दर्शाता है। रेड-लाइन मार्करों के बीच स्थायी संबंध है। सामान्य रैखिक मॉडल (GLM), मिश्रित रैखिक मॉडल (MLM) और संपीड़ित MLM (CMLM)।
यह अनुमानित है कि जीडब्ल्यूएएस जटिल लक्षणों के गुणात्मक अनुपात को पूरी तरह से समझाएगा, लेकिन, यह एक बड़े अनुपात की व्याख्या कर सकता है। दुर्लभ वेरिएंट द्वारा छोटे प्रभावों का पता लगाने में कठिनाई या सामान्य एलील द्वारा बहुत छोटे प्रभाव को असंभव बना देता है।

संदर्भ

[1] Alonso-Blanco C, Aarts MG, Bentsink L, Keurentjes JJ, Reymond M, Vreugdenhil D, et al. What has natural variation taught us about plant development, physiology, and adaptation?. Plant Cell 2009;21(7):1877–96.

[2] Mitchell-Olds T, Willis JH, Goldstein DB. Which evolutionary processes influence natural genetic variation for phenotypic traits?. Nat Rev Genet 2007;8(11):845–56.

[3] Pourkheirandish M, Hensel G, Kilian B, Senthil N, Chen G, Sameri M, et al. Evolution of the Grain Dispersal System in Barley. Cell 2015;162(3):527–39.

[4] Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, et al. Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. PNAS 2007;104(4):1424–9.

[5] Wambugu PW, Ndjiondjop MN, Henry RJ. Role of genomics in promoting the utilization of plant genetic resources in genebanks. Brief Funct Genomics. 2018;17(3):198–206.

[6] Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by- sequencing approach. PLoS ONE 2012;7(2):e32253.

[7] Kumar J, Pratap A, Solanki RK, Gupta DS, Goyal A, Chaturvedi SK, et al. Genomic resources for improving food legume crops. J Agric Sci 2012;150 (3):289–318.

[8] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904–9.

[9] Thabet SG, Moursi YS, Karam MA, Graner A, Alqudah AM. Genetic basis of drought tolerance during seed germination in barley. PLoS ONE 2018;13(11): e0206682.

[10] Flint-Garcia SA, Thornsberry JM. Buckler ESt. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 2003;54(1):357–74.

[11] Xu S. Theoretical basis of the Beavis effect. Genetics 2003;165(4):2259–68.

[12] Alqudah AM, Koppolu R, Wolde GM, Graner A, Schnurbusch T. The Genetic Architecture of Barley Plant Stature. Front Genet. 2016;7:117.

[13] Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 2007;23(19):2633–5

[14] Rentería ME, Cortes A, Medland SE. Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis. In: Gondro C, van der Werf J, Hayes B, editors. Genome-Wide Association Studies and Genomic Prediction. Totowa, NJ: Humana Press; 2013. p. 193–213

हाई डायमेंशनल बायोलॉजिकल डाटा एनालिसिस यूसिंग आर

डॉ. समरेन्द्र दास
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

सॉफ्टवेयर एक ओपन सोर्स सॉफ्टवेयर है और इसे http://CRAN.R-project.org से डाउनलोड किया जा सकता है। R सॉफ्टवेयर बहुत ही लचीले तरीके से यूनीवेरिएट और मल्टीवेरिएट विश्लेषण की सुविधा देता है। इसके अलावा, RStudio सॉफ्टवेयर सांख्यिकीय प्रोग्रामिंग सॉफ्टवेयर R. लर्निंग आर स्टूडियो के लिए एक उपयोगकर्ता इंटरफ़ेस है जो अंततः आर पर जैविक डेटा का विश्लेषण और कल्पना करते समय पूर्ण नियंत्रण, लचीलापन और रचनात्मकता देगा, लेकिन इस नई भाषा में प्रवाह में समय लगेगा। कंप्यूटर पर आर स्थापित करने के बाद, RStudio को http://www.rstudio.com/ से डाउनलोड किया जा सकता है, "अभी डाउनलोड करें" पर क्लिक करके और फिर "डाउनलोड RStudio डेस्कटॉप" पर क्लिक करें। पूर्ण डाउनलोड के बाद, "रन" पर क्लिक करके आर स्टूडियो स्थापित किया जा सकता है। इसके अलावा, आर स्टूडियो अधिक उपयोगकर्ता के अनुकूल है, सामान्य आर सॉफ़्टवेयर पर संचालित होता है और डेटा प्रबंधन और विज़ुअलाइज़ेशन के लिए कई और विकल्प हैं। इसमें कई उपयोगिताओं, पैकेज और फ़ंक्शन शामिल हैं जो डेटा विश्लेषण और ग्राफिक्स के लिए हैं, विशेष रूप से जीनोमिक डेटा के लिए। वर्तमान व्याख्यान में, आर / आर स्टूडियो सॉफ़्टवेयर का उपयोग करके जीनोमिक डेटा का विश्लेषण वर्णनात्मक आंकड़ों, सहसंबंध, प्रतिगमन, एनोवा, रैखिक मॉडल, प्रधान घटक विश्लेषण (पीसीए), क्लस्टर विश्लेषण, ग्राफ़, जीन अभिव्यक्ति विश्लेषण, जीन पर विशेष जोर देने के साथ किया जाता है। विनियमन मॉडलिंग, नेटवर्क विश्लेषण, डीएनए / आरएनए अनुक्रम विश्लेषण, आदि।

1. वर्णनात्मक आँकड़े

वर्णनात्मक आँकड़े डेटा में विशेष पैटर्न को समझने में मूल्यवान उपकरण हैं। इस अनुभाग के प्रयोजनों के लिए, हम मानेंगे कि आपके डेटा का उत्पादन करने वाले प्रयोग दो अलग-अलग डेटा प्रकारों में से एक का उत्पादन करते हैं। सबसे पहले, आपके डेटा से टिप्पणियों को यादृच्छिक चर माना जा सकता है; एक माप जो एक वास्तविक संख्या का उत्पादन करता है।

1.1 मूल आँकड़े

आर बुनियादी आंकड़ों के लिए कई कार्य प्रदान करता है। आंकड़ों के मूल पैटर्न को समझने के लिए एक ही संख्या पर कई प्रकार के कार्य किए जा सकते हैं। डेटा सेट में एक सरणी हो सकती है जिसे आपने .CSV या .txt फ़ाइल से पढ़ा है या यह बड़े डेटा सेट से हो सकता है। यदि बाद का मामला है, तो सुनिश्चित करें कि आप (डाटासेट) संलग्न करें ताकि बड़े डेटा सेट के भीतर निहित विभिन्न चर स्मृति में पढ़े जाएं।

The basic arithmetic mean	mean(variable)
The median (middle value)	median(variable)
The largest value in the variable	max(variable)
The smallest value in the variable	min(variable)
The standard deviation of the variable	sd(variable)
The number of items in the variable	length(variable)
The variance is given by this	var(variable)
Quantile using this function. Set the level to any value e.g. 0.25, 0.75 to return the appropriate quantile	quantile(variable, level)

संयोजन में इन और अन्य कार्यों का उपयोग करना संभव है जैसे कि आप कैलकुलेटर का उपयोग कर रहे थे। अन्य कार्यों में वर्गमूल निर्धारित करने के लिए sqrt (परिवर्तनशील) शामिल हैं। पावर फंक्शन उत्पन्न करने के लिए कैरेट कैरेक्टर का उपयोग करें उदा। 2 ^ 3 3 की शक्ति (यानी 8) को 2 देता है। यदि आपका डेटा सेट कई स्तंभों से बना है, तो वे सभी समान लंबाई के नहीं हो सकते हैं। डिफ़ॉल्ट रूप से एनए के साथ 'लापता' कोशिकाओं को बाहर निकालता है। यदि आपके चर में NA मान हैं तो यह आपकी गणनाओं को प्रभावित करेगा। इस उपयोग को पाने के लिए कमांड में na.rm = TRUE का उपयोग करें *e.g.* mean(variable, na.rm= TRUE).

2. परिकल्पना का परीक्षण

परिकल्पना प्रक्रियाओं के व्यापक परीक्षण के लिए आर कई कार्य प्रदान करता है; जिन्हें बड़े पैमाने पर जीनोमिक अनुसंधान के लिए उपयोग किया जाता है। इन टेस्ट में टी-टेस्ट, मैन-विटनी यू-टेस्ट, ची-स्क्वेर्ड टेस्ट, गुडनेस ऑफ फिट टेस्ट आदि शामिल हैं।

टी परीक्षण

टी-टेस्ट का उपयोग दो नमूनों के बीच सांख्यिकीय अंतर को निर्धारित करने के लिए किया जाता है। एक ऐसा संस्करण भी है जिसे एक युग्मित परीक्षण के रूप में उपयोग किया जा सकता है यानी जब आपके पास मिलान किए गए जोड़े के रूप में एकत्र किए गए माप होते हैं।

कदम:

(i) अपने डेटा को एक .CSV फ़ाइल में व्यवस्थित करें। प्रत्येक चर के लिए एक कॉलम का उपयोग करें और इसे एक सार्थक नाम दें। यह मत भूलो कि R में परिवर्तनशील नाम अक्षरों और संख्याओं को शामिल कर सकते हैं लेकिन केवल विराम चिह्न की अनुमति अवधि है।

(ii) अपनी डेटा फ़ाइल को मेमोरी में पढ़ें और इसे एक समझदार नाम दें।

(ii) अपने डेटा सेट को संलग्न करें ताकि व्यक्तिगत चर स्मृति में पढ़े जाएं।

एक टी-टेस्ट करने के लिए आप टाइप करें (उपयोग):

```
> t.test(var1, var2)
```

```
Example: data: x1 and x2
t = 4.0369, df = 22.343, p-value = 0.0005376
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2.238967 6.961033
sample estimates:
mean of x mean of y
8.733333 4.133333
```

परीक्षण का यह संस्करण यह नहीं मानता है कि दो नमूनों का विचरण समान है और एक वेल्च दो नमूना टी-परीक्षण करता है। टी-टेस्ट का "क्लासिक" संस्करण निम्नानुसार चलाया जा सकता है:

```
> t.test(var1, var2, var.equal=T)
```

```
Two Sample t-test
data: x1 and x2
t = 4.0369, df = 28, p-value = 0.0003806
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2.265883 6.934117
sample estimates:
mean of x mean of y
8.733333 4.133333
```

अब दो नमूनों के प्रकार को समान माना जाता है और मूल संस्करण का प्रदर्शन किया जाता है, जो वास्तविक जैविक डेटा के लिए बहुत दुर्लभ है। युग्मित डेटा पर एक टी-टेस्ट चलाने के लिए आप एक नया शब्द जोड़ते हैं:

```
> t.test(var1, var2, paired=T)
```

```
Paired t-test
data: x1 and x2
```

t = 4.3246, df = 14, p-value = 0.0006995
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2.318620 6.881380
sample estimates:
mean of the differences
4.6

Step by step procedure for t-test in R

Read in your file and assign it to a variable name	Your.data= read.csv(file.choose())
Make the variables within the data set available to R	attach(your.data)
For a classic t-test (variances assumed equal)	t.test(var1, var2, var.equal=T)
If variances are assumed unequal use the Welch procedure	t.test(var1, var2)
If you have paired data run the paired version	t.test(var1, var2, paired=T)

मान- विटनी U- टेस्ट

मान-विटनी U- परीक्षण आमतौर पर डेटा के गैर-पैरामीट्रिक होने पर दो नमूनों के बीच महत्वपूर्ण अंतर का परीक्षण करने के लिए उपयोग किया जाता है। आर में परीक्षण शायद भ्रामक रूप से विलकॉक्सन परीक्षण कहा जाता है और इसे दो नमूनों या युग्मित डेटा पर लागू किया जा सकता है।

आर में मान-विटनी यू-टेस्ट के लिए प्रक्रिया

पहला चरण एक .CSV फ़ाइल में आपके डेटा की व्यवस्था करना है। प्रत्येक चर के लिए एक कॉलम का उपयोग करें और इसे एक सार्थक नाम दें। यह मत भूलो कि R में परिवर्तनशील नाम अक्षरों और संख्याओं को शामिल कर सकते हैं लेकिन केवल विराम चिह्न की अनुमति अवधि है।

दूसरा चरण आपकी डेटा फ़ाइल को मेमोरी में पढ़ना और इसे एक समझदार नाम देना है।

अगला चरण आपके डेटा सेट को संलग्न करना है ताकि व्यक्तिगत चर को मेमोरी में पढ़ा जाए।

The basic u-test is performed on two samples so:

> wilcox.test(var1, var2)
Wilcoxon rank sum test with continuity correction
data: x1 and x3
W = 63.5, p-value = 0.04244
alternative hypothesis: true mu is not equal to 0
Warning message: cannot compute exact p-value with ties in: wilcox.test.default(x1, x3, paired = F)

If you have paired data you can run a matched pair test:

> wilcox.test(var1, var2, paired=T)
Wilcoxon signed rank test with continuity correction
data: x1 and x3
V = 22.5, p-value = 0.06299

alternative hypothesis: true mu is not equal to 0
Warning messages:
1: cannot compute exact p-value with ties in: wilcox.test.default(x1, x3, paired = T)
2: cannot compute exact p-value with zeroes in: wilcox.test.default(x1, x3, paired = T)

In the above examples we see that there are several warning messages. We can safely ignore these. Also, the test runs with continuity correction as the default. If you want to turn this off (I cannot see why you would) then add correct=F to the parameters e.g. > wilcox.test(var1, var2, correct=F)

यू-टेस्ट स्टेप बाय स्टेप
सबसे पहले अपनी डाटा फाइल बनाएं। स्प्रेडशीट का उपयोग करें और प्रत्येक कॉलम को एक चर बनाएं। प्रत्येक पंक्ति एक प्रतिकृति है लेकिन कॉलम में समान डेटा आइटम (जब तक आप एक युग्मित परीक्षण नहीं चाहते हैं) को समाहित करने की आवश्यकता नहीं है। पहली पंक्ति में चर नाम होने चाहिए। इसे .CSV फ़ाइल या .txt फ़ाइल के रूप में सहेजें

Read in your file and assign it to a variable name	Your.data= read.csv(file.choose())
Make the variables within the data set available to R	attach(your.data)
For a standard two-sample U-test	wilcox.test(var1, var2)
If you have paired data run the paired version	wilcox.test(var1, var2, paired=T)
The default tests run with continuity correction, to turn this off use	wilcox.test(var1, var2, correct=F)

Chi-squared tests
एसोसिएशन के लिए टेस्ट आसानी से आर में किए जाते हैं। बेसक फ़ंक्शन chisq.test () है। आर में ची-चुकता परीक्षण शामिल हैं:
(a) अपने डेटा को एक .CSV फ़ाइल में व्यवस्थित करें। पंक्ति और स्तंभ नामों का उपयोग करें। यह मत भूलो कि R में परिवर्तनशील नाम अक्षरों और संख्याओं को शामिल कर सकते हैं लेकिन केवल विराम चिह्न की अनुमति अवधि है।
(b) अपनी डेटा फ़ाइल को मेमोरी में पढ़ें और इसे एक समझदार नाम दें। आपको R को बताना होगा कि फ़ाइल में पंक्ति नाम हैं ताकि एक डेटा मैट्रिक्स बनाया जाए।
(c) Chi-स्क्वैयर परीक्षण करें जो आप कुछ इस प्रकार करते हैं:

```
chisq.test(your.data)
Pearson's Chi-squared test
data: your.data
X-squared = 121.5774, df = 8, p-value < 2.2e-16
```

यह आपको एक मूल परिणाम देता है, लेकिन आप सांख्यिकीय की व्याख्या करने के लिए इससे अधिक चाहते हैं।
परीक्षण से अधिक डेटा का उत्पादन होता है, यह देखने के लिए कि आपको किस प्रकार के साथ काम करना है:

```
> names(chisq.test(your.data))
[1] "statistic" "parameter" "p.value" "method" "data.name" "observed"
[7] "expected" "residuals"
```

Chi-Squared test in R (Step by Step)

Read in your file and assign it to a variable name.	Your.data=
This command tells R that the 1st column contains the row names.	read.csv(file.choose(), row.names=1)
Have a look at your data to see that it contains what you expected	your.data
Run the Chi-Squared test and assign it to a variable	your.chi= chisq.test(your.data)
If you need to apply Yates correction for a 2 x 2 matrix	your.chi = chisq.test(your.data, correct=T)
To see the original data i.e. observed values	your.chi$obs
To see the expected values	your.chi$exp
To see the Pearson residuals (O-E)/sqrt(E)	your.chi$res
To extract a single item from the Observed, Expected or Residual tables	your.chi$table["row", col"]

3. Correlation (सहसंबंध)

R, cor () फ़ंक्शन के साथ सहसंबंध कर सकता है। कार्यक्रम के आधार वितरण के लिए अंतर्निहित तीन मार्ग हैं; पियर्सन, केंडल और स्पीयरमैन रैंक के सहसंबंधों के लिए। प्रक्रिया में मुख्य रूप से शामिल हैं:

पहला चरण एक .CSV फ़ाइल में आपके डेटा की व्यवस्था करना है। प्रत्येक चर के लिए एक कॉलम का उपयोग करें और इसे एक सार्थक नाम दें। यह मत भूलो कि R में परिवर्तनशील नाम अक्षरों और संख्याओं को शामिल कर सकते हैं लेकिन केवल विराम चिह्न की अनुमति अवधि है।

दूसरा चरण आपकी डेटा फ़ाइल को मेमोरी में पढ़ना और इसे एक समझदार नाम देना है। अगला चरण आपके डेटा सेट को संलग्न करना है ताकि व्यक्तिगत चर को मेमोरी में पढ़ा जाए। सहसंबंध गुणांक प्राप्त करने के लिए आप टाइप करें:

> cor(var1, var2, method = "method")

डिफ़ॉल्ट विधि "पीयरसन" है, इसलिए यदि आप चाहते हैं तो आप इसे छोड़ सकते हैं। यदि आप "केंडल" या "स्पीयरमैन" टाइप करते हैं तो आपको उचित सहसंबंध गुणांक मिलेगा।

Correlation and Significance tests

सहसंबंध गुणांक प्राप्त करना आमतौर पर केवल आधी कहानी है; आप जानना चाहेंगे कि क्या संबंध महत्वपूर्ण है। cor() फ़ंक्शन को आर में बढ़ाया जा सकता है ताकि आवश्यक परीक्षण की आवश्यकता हो। फ़ंक्शन cor.test () है

जैसा कि ऊपर आपको अपने डेटा को एक .CSV फ़ाइल से R में पढ़ने और कारकों को संलग्न करने की आवश्यकता है ताकि वे सभी मेमोरी में संग्रहीत हों।

एक सहसंबंध परीक्षण चलाने के लिए हम टाइप करते हैं:

> cor.test(var1, var2, method = "method")

डिफ़ॉल्ट विधि "पीयरसन" है, इसलिए यदि आप चाहते हैं तो आप इसे छोड़ सकते हैं। यदि आप "केंडल" या "स्पीयरमैन" टाइप करते हैं तो आपको उपयुक्त महत्व परीक्षण मिलेगा।

सहसंबंध रेखांकन

आप आमतौर पर अपने सहसंबंध को रेखांकन करने के लिए स्कैटर प्लॉट का उपयोग करना चाहेंगे। मूल कथानक कथानक () है। आर में विभिन्न डिफ़ॉल्ट पैरामीटर सेट हैं, उदा। कुल्हाड़ियों को कारक नाम के रूप में लेबल किया जाता है और प्लॉटिंग प्रतीक को एक खुले सर्कल के रूप में सेट किया जाता है।

Correlation graphs

Use the basic defaults to create a	plot(x.var, y.var)

scatter plot of your two variables	
This changes the axes titles	plot(x.var, y.var, xlab="X-axis", ylab="Y-axis")
This changes the plotting symbol to a solid circle	plot(x.var, y.var, pch=16)
Adds a line of best fit to your scatter plot (don't do this for non-parametric plots).	abline(lm(y.var ~ x.var)

सहसंबंध विश्लेषण को निम्नलिखित तालिका में संक्षेपित किया जा सकता है:

Read the data into R and save as some name	your.data = read.csv(file.choose())
Allow the factors within the data to be accessible to R	attach(your.data)
Decide on the method, run the correlation and assign the result to a new variable. Methods are "pearson" (default), "kendal" and "spearman"	your.cor = cor(var1, var2, method = "pearson")
Have a look at the resulting correlation coefficient	your.cor
Perform a pairwize correlation on all the variables in the data set. Decide on the method ("pearson" (default), "kendal" and "spearman")	cor.mat = cor(your.data, method = "pearson")
have a look at the resulting correlation matrix	cor.mat
To evaluate the statistical significance of your correlation decide on the appropriate method (pearson is the default, see above), assign a variable and run the test	your.cor =cor.test(var1, var2, method="spearman")
Have a look at the result of yor significance test	your.cor
Plot a graph of the two variables from your correlation. pch=21 plots an open circle, pch=19 plots a solid circle. Try other values.	plot(x.var, y.var, xlab="x-label", ylab="y-label", pch=21))
Add a line of best fit (if appropriate)	abline(lm(y.var ~ x.var)

1. Multiple Regression Analysis

आर बहुत आसानी से कई प्रतिगमन विश्लेषण कर सकते हैं। मूल कार्य है: एलएम (मॉडल, डेटा)

पहला चरण एक .CSV फ़ाइल में आपके डेटा की व्यवस्था करना है। प्रत्येक चर के लिए एक कॉलम का उपयोग करें और इसे एक सार्थक नाम दें। यह मत भूलो कि R में परिवर्तनशील नाम अक्षरों और संख्याओं को शामिल कर सकते हैं लेकिन केवल विराम चिह्न की अनुमति अवधि है।

दूसरा चरण आपकी डेटा फ़ाइल को मेमोरी में पढ़ना और इसे एक समझदार नाम देना है।

अगला चरण आपके डेटा सेट को संलग्न करना है ताकि व्यक्तिगत चर को मेमोरी में पढ़ा जाए।

अंत में हमें मॉडल को परिभाषित करने और विश्लेषण चलाने की आवश्यकता है।

Linear Regression Models

एक रैखिक प्रतिगमन का मूल रूप है: y = m1x1 + m2x2 + m3x3 ... + c

Ys की एक श्रृंखला और X1, x2 आदि की एक श्रृंखला को देखते हुए हम गुणांक (एमएस) और इंटरसेप्ट (c) निर्धारित कर सकते हैं। हम कारकों की सापेक्ष शक्ति भी निर्धारित कर सकते हैं और प्रत्येक कारक (या संयोजन) कितनी अच्छी तरह सहसंबद्ध है।

R में हमारे मॉडल का सामान्य रूप है: y ~ X1 + x2 ...

आपके डेटा सेट के आधार पर कई विकल्प हैं। आइए उस स्थिति पर विचार करें जहां आपके पास एक आश्रित चर (y) और 3 कारक हैं जो आपको लगता है कि y का निर्धारण करने में महत्वपूर्ण हैं; हम उन्हें X1, x2 और x3 कहेंगे। वास्तव में हम उन्हें और अधिक सार्थक नाम देंगे। हम अपने मॉडल को कई तरीकों से सेट कर सकते हैं:

Model	Meaning
y ~ x1	y is modelled by x1 only, a simple regression
y ~ x1 + x2	y is modelled by x1 and x2 as in a multiple regression
y ~ x1 + x2 + x3	y is modelled by x1, x2 and x3 as in a multiple regression
y ~ x1 * x2	y is modelled by x1, x2 and also by the interaction between them

To run an analysis we use the lm() function on our data e.g.

```
> lm(y ~ x1 + x2 +x3)
```

हमें डेटा फ़ाइल को निर्दिष्ट करने की आवश्यकता नहीं है क्योंकि हमने पहले से ही इसे मेमोरी में पढ़ा है और वेरिएबल नामों को लिंक करने के लिए अटैच () का उपयोग किया है। विश्लेषण के परिणाम को "पकड़" करने के लिए एक चर का उपयोग करना अच्छा है; हम उस परिणाम पर अन्य काम कर सकते हैं जो हर बार टाइप करने के लिए थकाऊ होगा। इस उदाहरण में:

```
> field.lm = lm(y ~ x1 + x2 + x3)
```

यदि हम अब अपने नए वेरिएबल का नाम लिखते हैं, तो हमें परिणाम दिखाई देता है; कुछ इस तरह:

```
> field.lm

Call:

lm(formula = y ~ x1 + x2 + x3)

Coefficients:

  (Intercept)       x1       x2       x3

       4.8401   1.3196   0.8252   0.5266
```

यह ठीक है लेकिन जानकारी थोड़ी पतली है। थोड़ी और जानकारी प्राप्त करने के लिए हम सारांश (हमारे एलएम परिणाम) का उपयोग कर सकते हैं। इस मामले में हम देखेंगे:

```
> summary(field.lm)

Call:

lm(formula = y ~ x1 + x2 + x3)

Residuals:

     Min       1Q   Median       3Q      Max

-5.7190  -2.4540  -0.9873   2.9214   7.8078

Coefficients:

             Estimate   Std. Error   t value   Pr(>|t|)

(Intercept)  4.8401     8.9407       0.541     0.59906

x1           1.13196    0.3346       3.944     0.00230**

x2           0.8252     0.6320       1.306     0.21832

x3           0.5266     0.6999       0.752     0.46756

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

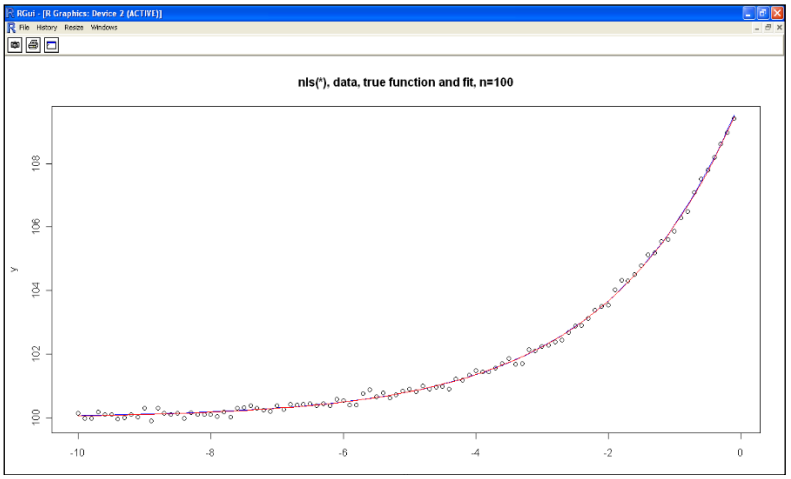
Residual standard error: 4.63 on 11 degrees of freedom
Multiple R-Squared: 0.6581, Adjusted R-squared: 0.5648
F-statistic: 7.056 on 3 and 11 DF, p-value: 0.0065

First create your data file. Use a spreadsheet and make each column a variable. Each row is a replicate. The first row should contain the variable names. Save this as a .CSV file	
Read the data into R and save as some name	your.data = read.csv(file.choose())
Allow the factors within the data to be accessible to R	attach(your.data)
Have a first look at the data as a pairs graph (plots all combinations as scatter plots)	pairs(your.data)
Decide on the model, run it and assign the result to a new variable	your.lm = lm(y.var ~ x1.var + x2.var + x3.var)
See the basic coefficients of your regression	your.lm
A more detailed summary of your regression	summary(your.lm)
Examine an individual coefficient	your.lm$coeff["x1.var"]
Calculate the beta coefficients (you will need to do one for each x factor)	beta.x1= your.lm$coeff["x1.var"] * sd(x1.var) / sd(y.var)
Display all your beta coefficients	cat(beta.x1, beta.x2, beta.x3)
Calculate the R-squared components (you will need to do one for each x factor)	R2.x1= beta.x1 * cor(y.var, x1.var)
Display all your R-squared values	cat(R2.x1, R2.x2, R2.x3)
Plot a graph of two variables from your regresion	plot(x.var, y.var, xlab="x-label", ylab="y-label"))
Add a line of best fit	abline(lm(y.var ~ x.var)

Non-linear regression model

गैर-रेखीय प्रतिगमन मॉडल को मानक फ़ंक्शन नेल्स () का उपयोग करके आर में फिट किया जा सकता है। एनएलएस फ़ंक्शन एक रिश्तेदार-ऑफसेट अभिसरण मानदंड का उपयोग करता है, जो वर्तमान पैरामीटर अनुमानों पर संख्यात्मक संसेचन की तुलना अवशिष्ट राशि-वर्ग में करता है। यह फॉर्म y = f (x, □) with (var (□)> 0) के डेटा पर अच्छा प्रदर्शन करता है। यह प्रपत्र y = f (x, because) के डेटा पर अभिसरण को इंगित करने में विफल रहता है क्योंकि मानदंड राउंड-ऑफ त्रुटि के दो घटकों की तुलना करने के लिए है। यदि कोई कृत्रिम डेटा पर nls का परीक्षण करना चाहता है, तो एक शोर जोड़ना होगा, जैसा कि नीचे दिखाया गया है।

```
x <- -(1:100)/10
y <- 100 + 10 * exp(x / 2) + rnorm(x)/10
nlmod <- nls(y ~  Const + A * exp(B * x), trace=TRUE)
plot(x,y, main = "nls(*), data, true function and fit, n=100")
curve(100 + 10 * exp(x / 2), col=4, add = TRUE)
lines(x, predict(nlmod), col=2)
```



>predict(nlmod)
>coef(nlmod)
>deviance(nlmod)
>vcov(nlmod)
>profile(nlmod)
>df.residual(nlmod)

Const A B
99.9870718 9.9547869 0.4966052

> deviance(nlmod)
[1] 1.078866

> vcov(nlmod)

	Const	A	B
Const	4.328019e-04	-1.552598e-05	8.228534e-05
A	-1.552598e-05	2.449755e-03	1.194106e-04
B	8.228534e-05	1.194106e-04	2.727130e-05

>profile(nlmod)
attr(,"original.fit")
Nonlinear regression model
model: y ~ Const + A * exp(B * x)
data: parent.frame()
Const A B
99.9870718 9.9547869 0.4966052
residual sum-of-squares: 1.078866

Number of iterations to convergence: 9
Achieved convergence tolerance: 1.379572e-07
attr(,"summary")
Formula: y ~ Const + A * exp(B * x)
Parameters:

	Estimate	Std. Error	t value	Pr(>|t|)	
Const	99.987071802	0.020803892	4806.17140	< 2.22e-16	***
A	9.954786890	0.049495004	201.12710	< 2.22e-16	***
B	0.496605232	0.005222193	95.09515	< 2.22e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1054625 on 97 degrees of freedom

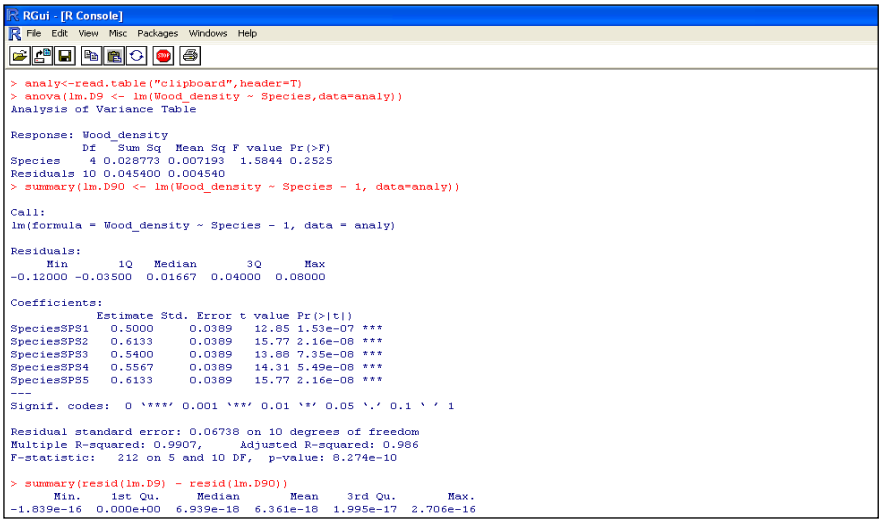
Number of iterations to convergence: 9
Achieved convergence tolerance: 1.379572e-07

5. Analysis of variance (ANOVA)

गन्ने की प्रजातियों के एक सेट से संबंधित एक यादृच्छिक रूप से एकत्र किए गए सेट पर प्राप्त लकड़ी के घनत्व पर टिप्पणियों के एक सेट पर विचार करें। प्रत्येक प्रजाति से आने वाली 3 टिप्पणियों के साथ 5 प्रजातियां होने दें। परिणाम नीचे दिए गए हैं।

Sl.No.	Species	Wood_density
1	SPS1	0.58
2	SPS1	0.54
3	SPS1	0.38
4	SPS2	0.53
5	SPS2	0.63
6	SPS2	0.68
7	SPS3	0.49
8	SPS3	0.55
9	SPS3	0.58
10	SPS4	0.53
11	SPS4	0.61
12	SPS4	0.53
13	SPS5	0.57
14	SPS5	0.64
15	SPS5	0.63

```
anova(lm.D9 <- lm(Wood_density ~ Species,data=analy))
summary(lm.D90 <- lm(weight ~ group - 1))# omitting intercept
summary(resid(lm.D9) - resid(lm.D90)) #- residuals almost identical
```



http://www.gardenersown.co.uk/education/lectures/r/correl.htm#correlation

अब तक हमने केवल एक सरल एक-तरफ़ा विश्लेषण पर विचार किया है। हालांकि, आपके पास अक्सर कई कारकों के साथ अधिक जटिल स्थिति होगी। कारकों के बीच बातचीत भी महत्वपूर्ण हो सकती है। सौभाग्य से आर में एक मॉडल वाक्यविन्यास है जो कई प्रकार के विश्लेषण के लिए काम करता है। एनोवा का संचालन करते समय हमारे पास एक एकल निर्भर चर और कई व्याख्यात्मक कारक होते हैं।

हमने एक सामान्य तरीके से अपनी एनोवा की स्थापना की: dependent ~ explanatory1... explanatory2...

मॉडल कई प्रकार के रूप ले सकता है:

Model	Meaning
y ~ x1	y is explained by x1 only, a one-way anova
y ~ x1 + x2	y is explained by x1 and x2, a two-way anova
y ~ x1 + x2 + x3	y is explained by x1, x2 and x3, a 3-way anova
y ~ x1 * x2	y is explained by x1, x2 and also by the interaction between them

Further, the step by step procedure for ANOVA in R can be well depicted in following table:

ANOVA Step by Step

सबसे पहले अपनी डाटा फाइल बनाएं। स्प्रेडशीट का उपयोग करें और प्रत्येक कॉलम को एक चर बनाएं। प्रत्येक पंक्ति एक प्रतिकृति है। पहली पंक्ति में चर नाम होने चाहिए। इसे .CSV फ़ाइल के रूप में सहेजें |

Read your data into R and assign a variable to it. This opens up a window and you select your file.	your.data = read.csv(file.choose())
Allow R to read the variables within the data file.	attach(your.data)
Decide on the anova model and run the analysis	your.aov = aov(dependent ~ explanatory)
View the result	summary(your.aov)
Carry out pairwise post-hoc testing using Tukey HSD test	TukeyHSD(your.aov)

Experimental Designs

(A) Completely Randomized Design (CRD): The following commands are used to analyze data obtained from CRD.

>crd<-read.table("clipboard",header=T)
>summary(fm1<-aov(Cao_gms ~ pullet, data=crd)) #Cao_gms is response variable and
>pullet is the classification variable
>TukeyHSD(fm1, "pullet", ordered = TRUE)
>plot(TukeyHSD(fm1, "pullet"))

(B)Randomized Complete Block Design (RCBD): Data obtained from experiments laid out in RCBD can be analyzed by the following commands.

>rbd<-read.table("clipboard",header=T)
>summary(fm1<-lm(yld ~ trt + block, data=rbd))
>TukeyHSD(fm1, "trt", ordered = TRUE)
>plot(TukeyHSD(fm1, "trt"))

(C) Latin Square Designs (LSD): Experimental data obtained from LSD is analyzed in the following way in R.

>lsd<-read.table("clipboard",header=T)
>summary(fm2<-aov(count ~ row+col+trt, data=lsd))
>TukeyHSD(fm1, "trt", ordered = TRUE)
>plot(TukeyHSD(fm1, "trt"))

7. Principal Components Analysis (PCA) and Cluster Analysis

>wea1<-read.table("clipboard",header=T) # Code meant for PCA analysis
>summary(wea1)
>print(summary(princomp(wea1, cor=F),loadings=T,cutoff=0.0001)) or
>res2=princomp(wea1,cor=F)
>print(summary(res2),cutoff=0.001)
>print(loadings(res2),cutoff=0.001)
>plot(res2,main="")
>library(MASS)
>eqscplot(res2$scores[,1:2],type="n",xlab="First Principal Component",ylab="second principal component")
>text(res2$scores[,1:2],labels=row.names(wea1))
>biplot(res2)
>diswea1=dist(wea1,method="euclidean") # Code meant for Cluster analysis
>hclustwea1=hclust(diswea1,method="average")
>par(mfrow=c(2,1),mar=c(0,4,0,0))
>plclust(hclustwea1,sub="",xlab="")
>ctreeclus=cutree(hclustwea1,k=3)
>options(width=68)
>carclusternames=sapply(1:3,function(nc)row.names(wea1)[ctreeclus==nc])
>names(carclusternames)=paste("cluster",1:3,sep="") carclustername

>carclusters=lapply(1:3,function(nc) wea1[ctreeclus==nc,])
>names(carclusters)=paste("cluster",1:3,sep="")
>carclusterss
>carclusmean=sapply(1:3,function(nc)apply(wea1[ctreeclus==nc,],2,mean))
>colnames(carclusmean)=paste("cluster",1:3,sep="")
>carclusmean

Genomic data analysis using R

बायोकॉन्टर (Bioconductor) पैकेज रिपॉजिटरी के साथ आर सॉफ्टवेयर उच्च-थ्रूपुट अनुक्रम डेटा के विश्लेषण के लिए बहुत लोकप्रिय है। संबद्ध बायोकॉन्टर प्रोजेक्ट विभिन्न जीवन विज्ञान क्षेत्रों में सांख्यिकीय डेटा विश्लेषण के लिए कई अतिरिक्त आर पैकेज प्रदान करता है, जैसे कि माइक्रोएरे, अगली पीढ़ी के अनुक्रम और जीनोम विश्लेषण के लिए उपकरण। बायोकॉन्टर आर सांख्यिकीय प्रोग्रामिंग भाषा का उपयोग करता है और खुला स्रोत और खुला विकास है। यह हर साल दो रिलीज, 1296 सॉफ्टवेयर पैकेज और एक सक्रिय उपयोगकर्ता समुदाय है। बायोकॉन्टर को आर प्लेटफॉर्म पर टाइप करके स्थापित किया जा सकता है:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite()
```

Biological sequence analysis

प्रायिकता और संभाव्यता वितरण का सिद्धांत ज्यादातर जीनोम अनुक्रम विश्लेषण के लिए उपयोग किया जाता है। संभावना एक प्रयोग के यादृच्छिक परिणामों को देखने पर आधारित है। गणितीय परिणामों का उपयोग करके इन परिणामों को मॉडल करने के लिए, हम "यादृच्छिक चर" नामक चर का उपयोग करते हैं। यादृच्छिक चर एक प्रयोग के प्रत्येक परिणाम के लिए एक संख्यात्मक मूल्य प्रदान करते हैं।
उदाहरण के लिए, नीचे RNA अनुक्रम पर विचार करें:
AUGCUUCGAAUGCUGUAUGAUGUC
इस क्रम में 5 A, 9 U, 6 G और C सी कुल 24 अवशेष हैं। इस क्रम को मॉडल करने के लिए, यादृच्छिक चर एक्स का उपयोग किया जा सकता है जहां एक्स न्यूक्लियोटाइड अवशेषों का प्रतिनिधित्व करता है। क्योंकि मात्रात्मक जानकारी के साथ काम करने के फायदे हैं, जब डेटा को गुणात्मक रूप से वर्णित किया जाता है तो एक संख्या को गैर-संख्यात्मक परिणामों के लिए असाइन करने के लिए एक यादृच्छिक चर का उपयोग किया जाता है। इस प्रयोग के लिए, A को 0, C को 1, G के रूप में 2 और U के रूप में 3 के रूप में प्रतिनिधित्व करने वाले यादृच्छिक चर मान असाइन करें। एक छोटा अक्षर यादृच्छिक चर के परिणाम का प्रतिनिधित्व करता है, इसलिए यहां छोटे x का उपयोग किया जा सकता है। इसलिए, प्रायिकता के संदर्भ में, इस प्रयोग के लिए यादृच्छिक चर X का उपयोग करते हुए मॉडल निम्न तालिका में दिया गया है।

Using Random Variable X to Quantitatively Model Residues in a Particular RNA Sequence

Residue	Value of X (=x)	P (X=x)
A	0	5/24=0.208
C	1	4/24=0.167
G	2	6/24=0.25
U	3	9/24=0.375

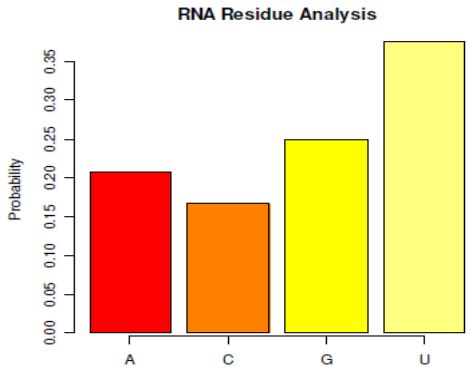
यदि प्रयोग आरएनए के एक अन्य अनुक्रम में प्रत्येक न्यूक्लियोटाइड की आवृत्ति की गणना करने के लिए है, तो यादृच्छिक चर के मान समान

होंगे लेकिन यह मानने वाले यादृच्छिक चर की संभावनाएं हैं कि प्रयोग के विभिन्न परीक्षणों को दर्शाते हुए मूल्य थोड़ा भिन्न होगा। इस सरल मॉडल (जो एक समीकरण का उपयोग भी नहीं करता है) को समझना अधिक जटिल मॉडल को समझने की कुंजी है। संभावना मॉडल बस प्रयोग के परिणाम को दर्शाने के लिए यादृच्छिक चर का उपयोग करते हैं चाहे वह एक साधारण प्रयोग हो (जैसा कि ऊपर है) या कई परिणामों के साथ बहुत अधिक जटिल प्रयोग।

हर रैंडम वैरिएबल में एक सम्भावना प्रायिकता वितरण फंक्शन होता है। इस फ़ंक्शन को असतत रैंडम वैरिएबल के मामले में संभाव्यता मास फंक्शन या एक सतत रैंडम वैरिएबल के मामले में प्रायिकता घनत्व फ़ंक्शन कहा जाता है। वितरण फ़ंक्शन का उपयोग यह देखने के लिए किया जाता है कि परिणाम संभावनाएं यादृच्छिक चर के मूल्यों से कैसे जुड़ी हैं। इसके अलावा सभी यादृच्छिक चर (असतत और निरंतर) में एक संचयी वितरण फ़ंक्शन, या CDF है। CDF एक ऐसा फंक्शन है जो यह संभावना देता है कि रैंडम वेरिएबल X हर वैल्यू x के लिए x के बराबर या उससे कम है, और उस वैल्यू तक संचित प्रायिकता को मॉडल करता है।

आर में एक साधारण हिस्टोग्राम (नीचे चित्रा में शो) का उपयोग इस उदाहरण के लिए संभाव्यता वितरण फ़ंक्शन को मॉडल करने के लिए किया जा सकता है।

```
> X<-c(0,1,2,3)
> Prob<-c(0.208,0.167,0.25,0.375)
> N<-c (‘A’, ‘C’, ‘G’, ‘U’)
> barplot(Prob,names=N,ylab="Probability", main="RNA Residue Analysis")
```



इस उदाहरण के लिए संचयी वितरण मान को खोजने के लिए, बस X के प्रत्येक मान के लिए 0,1,2,3 के लिए संभाव्यताएं जोड़ें और CDF का मान वह संभावना है जो यादृच्छिक चर X मानता है या उससे कम मूल्य का है। उदाहरण के लिए यदि X 2 के बराबर है, तो CDF संभावना है कि X = 2 या X = 1 या X = 0। इसे बस पी (एक्स = 2) प्लस पी (एक्स = 1) प्लस पी (एक्स = 0) के लिए मानों की गणना करने के लिए। हमारे आरएनए अवशेष उदाहरण के लिए, सीडीएफ के लिए गणना नीचे तालिका में दिखाई गई है।

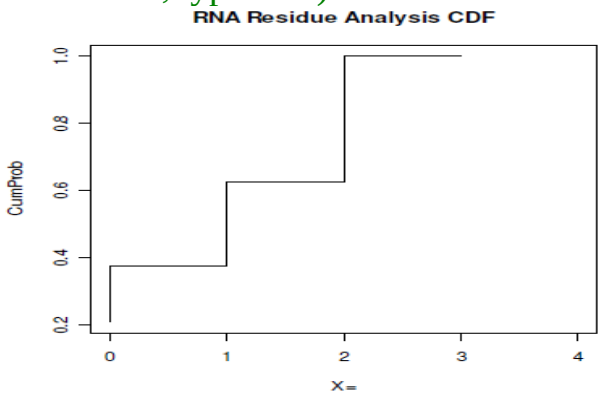
Probability Distribution and Cumulative Probability Distribution for RNA Residue Analysis

Residue	Value of X (=x)	P (X=x)	F(x)= P(X≤x)
A	0	5/24=0.208	0.208
C	1	4/24=0.167	0.375
G	2	6/24=0.25	0.625
U	3	9/24=0.375	1

सीडीएफ का नेत्रहीन विश्लेषण करने के लिए, इस सीडीएफ का एक सरल चरण ग्राफ पिछले कोड के नीचे कमांड जोड़कर आर में किया जा सकता है। नीचे दिया गया चित्र CDF के प्लॉट को निम्न कोड द्वारा निर्मित दिखाता है।

```
> CumProb<-c(0.208, 0.375, 0.625, 1)
```

```
> plot(X,CumProb,xlim=range(0,1,2,3,4), main="RNA Residue Analysis CDF",
xlab="X=", type="S")
```



Expression data analysis

जीन अभिव्यक्ति विश्लेषण के लिए अधिकांश अत्याधुनिक तरीके आर प्रणाली के लिए पैकेज के रूप में भी उपलब्ध हैं (और कभी-कभी विशेष रूप से भी)। आर एक ओपन-सोर्स है और व्यापक रूप से इस्तेमाल किया जाने वाला और बहुत ही बहुमुखी सांख्यिकी पैकेज है। आर विंडोज़, मैकिन्टोश और लिनक्स / यूनिक्स सिस्टम के लिए स्वतंत्र रूप से उपलब्ध है। इसलिए, आर के साथ जीन अभिव्यक्ति डेटा का विश्लेषण करने के लिए दृढ़ता से अनुशंसा की जाती है। जीन अभिव्यक्ति विश्लेषण के लिए आर पैकेजों की एक काफी विस्तृत सूची उपलब्ध है।

इन पैकेजों में से एक विशेष रूप से उल्लेख के योग्य है: बायोकॉन्टर प्रोजेक्ट (हार्वर्ड) कई पूर्व स्वतंत्र आर पैकेजों को एकीकृत और विलय करता है और दोनों सीडीएनए, एफिमेट्रिक्स सरणियों, आरएनए-सिक डेटा के विश्लेषण के लिए उपकरण प्रदान करता है।

उदाहरण के लिए: हमने जीन अभिव्यक्ति डेटा लिया है जिसमें अभिगम संख्या GSE14403 है, जिसमें वनस्पति विकास के दौरान नियंत्रण और लवणता-तनावग्रस्त परिस्थितियों के 23 नमूनों के साथ नमक-सहिष्णु जीनोटाइप FL478, पोक्कली और IR63731 और नमक-संवेदनशील जीनोटाइप IR29 की जड़ जीन अभिव्यक्ति का विश्लेषण शामिल है। (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14403 पर उपलब्ध है)। मूल रूप से डेटा .CEL फ़ाइलों में उपलब्ध है, जिसका उपयोग सीधे नहीं किया जा सकता है। तो, आगे की सांख्यिकीय विश्लेषण से पहले कच्ची सीएल फ़ाइलों को पूर्व-संसाधित करने की आवश्यकता है और आगे, आर बायोकॉन्टर के साथ इस तरह के डेटा के लिए सांख्यिकीय विश्लेषण के लिए उत्कृष्ट मंच प्रदान करता है।

Affy पैकेज स्थापित करने के लिए (क्योंकि यह कच्चे cel फ़ाइल डेटा के पूर्व-प्रसंस्करण के लिए बड़े पैमाने पर उपयोग किया जाता है), आर शुरू करें और दर्ज करें:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("affy")
```

यदि आप चाहते हैं कि जांच के स्तर के डेटा (सेल फाइलें) से अभिव्यक्ति के उपायों तक जाना यहां कुछ त्वरित तरीके हैं। यदि आप आरएमए चाहते हैं, तो डेटा को पढ़ने और अभिव्यक्ति के उपायों को प्राप्त करने का सबसे तेज़ तरीका निम्नलिखित है:

1. एक निर्देशिका बनाएँ, सभी प्रासंगिक CEL फ़ाइलों को उस निर्देशिका में स्थानांतरित करें।
2. यदि Linux / Unix का उपयोग कर रहे हैं, तो उस निर्देशिका में आर शुरू करें।
3. यदि Microsoft Windows के लिए Rgui का उपयोग करते हैं, तो सुनिश्चित करें कि आपकी कार्यशील निर्देशिका में Cel फ़ाइलें हैं ("फ़ाइल -> परिवर्तन डार" मेनू आइटम का उपयोग करें)।
4. Library लोड करें।

> library(affy) ##load the affy package.

5. डेटा में पढ़ें और उदाहरण के लिए RMA का उपयोग करके एक अभिव्यक्ति बनाएं।

> Data <- ReadAffy() ##read data in working directory
> eset <- rma(Data)

आपके डेटासेट के आकार और आपके सिस्टम के लिए उपलब्ध मेमोरी के आधार पर, आपको त्रुटियों का अनुभव हो सकता है जैसे। वेक्टर को आवंटित नहीं किया जा सकता है ... । एक स्पष्ट विकल्प आपके आर प्रक्रिया के लिए उपलब्ध मेमोरी को बढ़ाना है (मेमोरी को जोड़ने और / या बाहरी अनुप्रयोगों को बंद करके। एक अन्य विकल्प फ़ंक्शन justRMA () का उपयोग करना है।

> eset <- justRMA()
> ExpressionSet <- exprs(eset)

यह डेटा को पढ़ता है और सी स्तर पर उन्हें प्रीप्रोसेस करने का ds RMA 'तरीका करता है। ReadAffy को कॉल करने की आवश्यकता नहीं है, जांच स्तर का डेटा कभी भी AffyBatch में संग्रहीत नहीं किया जाता है। RMA RMA की गणना के लिए अनुशंसित कार्य जारी है। Rma फ़ंक्शन गति और दक्षता के लिए C में लिखा गया था। इसके अलावा, द्रव्यमान जैसी विधि का उपयोग rma के बजाय एक्सप्रेसो में किया जा सकता है। उदाहरण के लिए MAS 5.0 सिग्नल के हमारे संस्करण के लिए एक्सप्रेस (कोड देखें) का उपयोग किया जाता है। 5.0 प्राप्त करने के लिए आप R> का उपयोग कर सकते हैं R> eset <- mas5(Data) which will also normalize the expression values.

उपरोक्त सभी उदाहरणों में, वैरिएबल एसेट क्लास एक्सप्रेशनसेट का एक ऑब्जेक्ट है जिसे बायोबेस विगनेट में वर्णित किया गया है। बायोकॉन्टर के कई पैकेज इस वर्ग की वस्तुओं पर काम करते हैं। कुछ उदाहरणों के लिए गेनफिल्टर और जीनप्लटर पैकेज देखें। यदि आप कुछ अन्य विश्लेषण पैकेज का उपयोग करना चाहते हैं, तो आप निम्न कमांड का उपयोग करके फाइल करने के लिए अभिव्यक्ति मान लिख सकते हैं:

R> write.exprs(eset, file="mydata.txt")

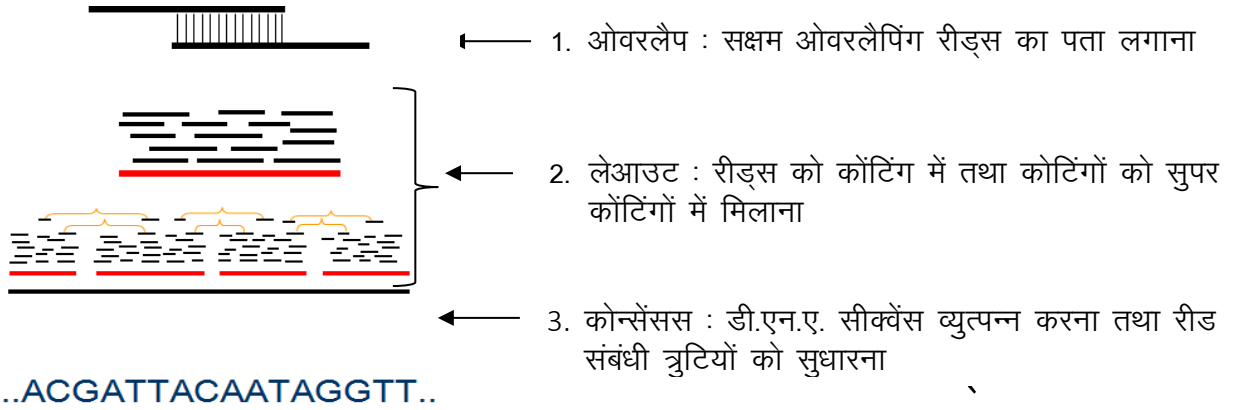
जीनोम समवेतीकरण: संकल्पना एवं चुनौतियाँ

डॉ. डी. सी. मिश्रा
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

प्रस्तावना

सीक्वेंस असेम्बली का अर्थ मूल सीक्वेंस के निर्माण के लिए डी.एन.ए. सीक्वेंस के खंडों को एलाइन करना व उन्हें आपस में मिलाना है। यह अपरिहार्य है क्योंकि डी.एन.ए. सीक्वेंसिंग प्रौद्योगिकी से एक बार में सम्पूर्ण जीनोमों को नहीं पढ़ा जा सकता है बल्कि उन्हें खंडों में पढ़ा जाता है और बेतरतीब क्रम में पढ़ा जाता है जो 20 और 1000 बेसों के बीच होते हैं और यह प्रयुक्त प्रौद्योगिकी पर निर्भर करता है। सीक्वेंसिंग प्रौद्योगिकी में हुई हाल की प्रगतियों से बड़ी मात्रा में सीक्वेंस आंकड़े सृजित करना संभव हुआ है। इन उच्च – थ्रू पुट विधियों से उत्पन्न खंड, तथापि, परंपरागत सेंगर सीक्वेंसिंग विधि की तुलना में काफी छोटे होते हैं।

प्रथम सीक्वेंस एसेम्बलर 1980 के दशक के अंत में तथा 1990 के दशक के आरंभ में साधारण सीक्वेंस एलाइनमेंट कार्यक्रमों के वेरिएंट के रूप में देखे गए थे जिनमें डी.एन.ए. सीक्वेंसर कहलाने वाले स्वचालित सीक्वेंसिंग इंस्ट्रूमेंट्स द्वारा बड़ी मात्रा में खंड सृजित हुए थे। सम्पूर्ण जीनोम शॉटगन (WGS) खंड असेम्बली के लिए एल्गोरिथ्म विकसित किए गए जिनमें एटलस, एराक्ने, सेलेरा, पीसीएपी, फ्रैप (www.phrap.org) और फ्यूजन शामिल हैं। ये सभी कार्यक्रम ओवरलैप–ले आउट – कन्सेंसस एप्रोच पर आधारित होते हैं जहां सभी रीड्स की तुलना युग्मवार फैशन में एक–दूसरे से की जाती है।



परिणामस्वरूप प्राप्त (ड्राफ्ट) जीनोम सीक्वेंस क्रमबद्ध किए गए 'कोंटिंग्स' की सूचना को मिलाकर किया जाता है और उसके बाद 'स्कैफोल्ड' सृजित करने के लिए संबंधित सूचना का उपयोग किया जाता है (चित्र 1)। स्कैफोल्ड 'सुनहरा पथ' सृजित करने के लिए गुणसूत्रों के भौतिक मानचित्र के साथ स्थित होते हैं।

हाल ही में एक नई अनुक्रमण विधि विकसित हुई है। वाणिज्यिक रूप से उपलब्ध प्रौद्योगिकियों में शामिल हैं : पाइरोसीक्वेंसिंग (454 सीक्वेंसिंग), संश्लेषण द्वारा सीक्वेंसिंग (इल्यूमिना) और लाइगेशन द्वारा सीक्वेंसिंग (एसओएलआईडी)। इन अगली पीढ़ी की सीक्वेंसिंग प्रौद्योगिकियों द्वारा सृजित रीड्स परंपरागत सैंगर रीड्स की तुलना में काफी छोटे होते हैं। अपनी छोटी लंबाई के कारण इन्हें बड़ी मात्रा में उत्पन्न किया जाना चाहिए तथा पूर्व की सीक्वेंसिंग तकनीकों की तुलना में इनके द्वारा अधिक सीक्वेंसिंग डेफ्थ होनी चाहिए जबकि लंबे रीड्स से लंबे ओवरलैप उपलब्ध होते हैं जिनसे वास्तविक ओवरलैप सुस्पष्ट हो जाते हैं, रिपीट्स में छोटे रीडों के अंतर निर्धारित किए जा सकते हैं। इन मुद्दों के कारण इन अत्यंत छोटे रीड्स के लिए विशेष रूप से नई असेम्बली युक्तियां डिजाइन करने के लिए

अनेक अनुसंधान दल उभर कर सामने आए हैं।

सीक्वेंसर के प्रकार और डेटा फार्मेट

इल्यूमिना	:	FASTQ
SoLID/ABI-Life	:	FASTA
Roche 454	:	SFF
Ion Torrent	:	SFF या FASTQ

असेम्बली के प्रकार

संदर्भ जीनोम की उपलब्धता के आधार पर असेम्बली के दो प्रकार हैं :

क) डी नोवो असेम्बली : रीड्स एक–दूसरे के साथ एलाइन किए जाते हैं ताकि कन्सेंसस सीक्वेंस निर्मित हो सके जो कोंटिंग कहलाते हैं।

ख) संदर्भ जीनोम एसेम्बली : यहां रीड्स एक कन्सेंसस सीक्वेंस का निर्माण करने के लिए उपलब्ध संदर्भ जीनोम से एलाइन किए जाते हैं।

जीनोम असेम्बली की तकनीकें

लगभग सभी बड़े पैमाने की सीक्वेंसिंग परियोजनाओं में ऐसी शॉटगन कार्यनीति का उपयोग होता है जिसमें एसेम्बलर (डेड्यूस) लक्षित क्रम से छोटे डी.एन.ए. खंडों के सैट से डी.एन.ए. सीक्वेंस को लक्षित करता है। छोटे डी.एन.ए. खंडों का सैट जो शॉटगन रीड्स कहलाता है, कोंटिंग या एलाइंड खण्डों के सैट के रूप में असेम्बल किया जाता है जिसके लिए फ्रेग्मेंट असेम्बलर नामक कलन विधि का उपयोग होता है। फ्रेग्मेंट असेम्बली एक संकल्पनात्मक सरल प्रक्रिया है जिसमें ओवरलैपिंग खण्डों की पहचान करके अपेक्षाकृत लंबे सीक्वेंस सृजित किए जाते हैं। यदि खण्ड एसेम्बली को सटीकता से किया जाए तो जीनोम सीक्वेंसिंग की समस्या सरल हो जाती है। तथापि, पुनरावृत्ति सीक्वेंस होते हैं जो अल्पकाल में पुनरावृत्ति होते हैं तथा जीनोमी क्रम में रहते हैं जिनसे खण्ड असेम्बली की प्रक्रिया में बहुत आसानी से गलती हो सकती है। रिपीट्स से उत्पन्न होने वाली कठिनाई को दूर करने के लिए उपयोगी तकनीक यह है कि क्लोन के दोनों छोरों को सीक्वेंस किया जाए जिससे प्रतिक्लोन दो खण्ड रीड सृजित हों। चूंकि क्लोन का इन्सर्ट आकार ज्ञात होता है अतः हम दो खंडों के बीच की लगभग दूरी जानते हैं। खण्ड मिलान संबंधी सूचना भी अक्सर मेट–पेयर सूचना कही जाती है जो बड़े पैमाने पर शॉटगन सीक्वेंसिंग के लिए अनिवार्य हो जाती है। असेम्बली प्रक्रिया के दौरान इस सूचना के उपयोग में मुख्य मुद्दा यह है कि हम दो रीड्स के बीच के क्रम को नहीं जानते हैं और इसे केवल एकल कोंटिंग में अन्य खंडों की असेम्बली द्वारा ही ज्ञात किया जा सकता है। इसलिए हम क्लोन – लंबाई संबंधी सूचना का उपयोग केवल असेम्बली के पश्चात कर सकते हैं जिससे क्लोन– लंबाई की सूचना के आधार पर सही या गलत, दोनों प्रकार की असेम्बली हो सकती है। मेट–पेयर सूचना के प्रभावी उपयोग की एक कार्यनीति सक्षम मिसअसेम्बल कोंटिंग्स का पता लगाकर यथासंभव सटीक कोंटिंग्स को असेम्बल करना तथा इसके बाद सही रूप से असेम्बल किए गए कोंटिंग्स का ही उपयोग करके मेट–पेयर सूचना का इस्तेमाल करना है। जीनोम–सीक्वेंसिंग केन्द्रों में जीनोम सीक्वेंसिंग तथा असेम्बली पर बल देने के लिए इस्तेमाल होने वाली सामान्य प्रक्रिया है :

1. खण्ड रीडआउटः प्रत्येक खण्ड का सीक्वेंस स्वचालित बेस–कालिंग सॉफ्टवेयर का उपयोग करके पता लगाया जाता है। फ्रैड सर्वाधिक व्यापक रूप से प्रयुक्त होने वाला कलन विधि है।

2. वाहक सीक्वेंस को कतरनाः शॉटगन रीड्स में वाहक क्रमों का वह भाग होता है जिसे सीक्वेंस असेम्बली के पूर्व हटाना होता है।
3. निम्न गुणवत्ता वाले क्रमों को कतरनाः शॉटगन रीड्स में घटिया गुणवत्ता वाले बेस काल होते हैं और इन निम्न गुणवत्ता वाले बेस काल को हटाने या उन्हें ढक देने से अक्सर अधिक सटीक सीक्वेंस असेम्बली होती है। तथापि यह चरण वैकल्पिक है तथा सीक्वेंसिंग करने वाले कुछ केन्द्र निम्न गुणवत्ता वाले बेस कालों को ढकते नहीं हैं तथा सच्चे खण्ड ओवरलैपों के बारे में निर्णय लेने के लिए गुणवत्तापूर्ण मानों के उपयोग की दृष्टि से फ्रेगमेंट असेम्बलर पर निर्भर रहते हैं।
4. खण्ड असेम्बलीः शॉटगन डेटा उस खण्ड असेम्बलर के लिए इनपुट है जो कोंटिंग्स कहलाने वाले एलाइन किए गए खण्ड के सैट को स्वतः ही सृजित करता है।
5. असेम्बली सत्यापनः पिछले चरणों में असेम्बल किए गए कुछ कोंटिंग रिपीट के कारण मिसअसेम्बल हो जाते हैं। चूंकि हमें लक्ष्य डी.एन.ए. में रिपीटों की पूर्व जानकारी नहीं होती है अतः प्रत्येक कोंटिंग में असेम्बली के सही होने को सत्यापित करना बहुत कठिन है और यह चरण अधिकांशतः मानवीय विधि से सम्पन्न किया जाता है। कोंटिंग असेम्बलियों के स्वचालित सत्यापन से संबंधित हाल ही में कुछ एल्गोरिद्मिक विकास हुए हैं।
6. स्कैफोल्डिंग कोंटिगः कोंटिंग अभिमुख तथा क्रमबद्ध होने चाहिए। मेट–पेयर सूचना इस चरण के लिए प्राथमिक सूचना है, अतः यदि इनपुट शॉटगन को क्लोनों के दोनों छोरों की रीडिंग द्वारा तैयार नहीं किया जाता है तो यह चरण पूरा नहीं हो सकता है।
7. समाप्तिः यह ज्ञात करने के लिए कि सभी कोंटिंग ठीक से असेम्बल हुए हैं और कोंटिंग अभिमुखित हैं व सही क्रम में हैं, हम अंतरालों की स्थिति से सम्बद्ध सीक्वेंसिंग विशिष्ट क्षेत्रों द्वारा दो कोंटिंग के बीच के अंतराल को भर सकते हैं।

अगली पीढ़ी के सीक्वेंसिंग (एनजीएस) रीड्स की नवीन असेम्बली

एनजीएस रीड सृजित होने के पश्चात् उन्हें ज्ञात संदर्भ सीक्वेंस के रूप में एलाइन किया जाता है या नवीन रूप से असेम्बल किया जाता है। नवीन असेम्बली जीवों के जीनोम की पुनर्संरचना की प्रक्रिया है। ये जीव इससे पहले सीक्वेंस नहीं किए गए होते हैं या इनके संदर्भ तुलनात्मक जीनोम उपलब्ध नहीं होते हैं। इसे शॉटगन प्रक्रिया से सम्पन्न किया जाता है जहां जीव के जीनोम को छोटे खण्डों में विभाजित किया जाता है और प्रत्येक को अलग–अलग सीक्वेंस किया जाता है या कम्प्यूटेशनल युक्तियों का उपयेाग करके उन्हें पुनः निर्मित किया जाता है। यह प्रक्रिया जटिल है क्योंकि जीनोम में समरूप सीक्वेंस के खण्ड होते हैं जो रिपीट कहलाते हैं। रिपीटों की लंबाई अत्यधिक भिन्न होती है जिससे सम्पूर्ण जीनोम को प्राप्त करना असंभव हो जाता है। इसलिए, लगभग सभी नई युक्तियों से पूर्ण जीनोम को प्राप्त नहीं किया जा सकता है। तथापि, इनमें कोंटिंग के नाम से ज्ञात जीनोम के लंबे खण्ड होते हैं। इसके अतिरिक्त जीनोम के आकार के बढ़ने के साथ जटिलता भी बढ़ती जाती है। नवीन जीनोम असेम्बली की प्रक्रिया में प्राथमिकतः दो श्रेणियां होती हैं, नामतः ओवरलैप लेआउट एंड कोंसेंसस (ओएलसी) तथा डी ब्रूजिन ग्राफ आधारित विधि। इनमें से पहली विधि मैमोरी गहन है। डी ब्रूजिन ग्राफ पर आधारित अनेक युक्तियां उपलब्ध हैं।

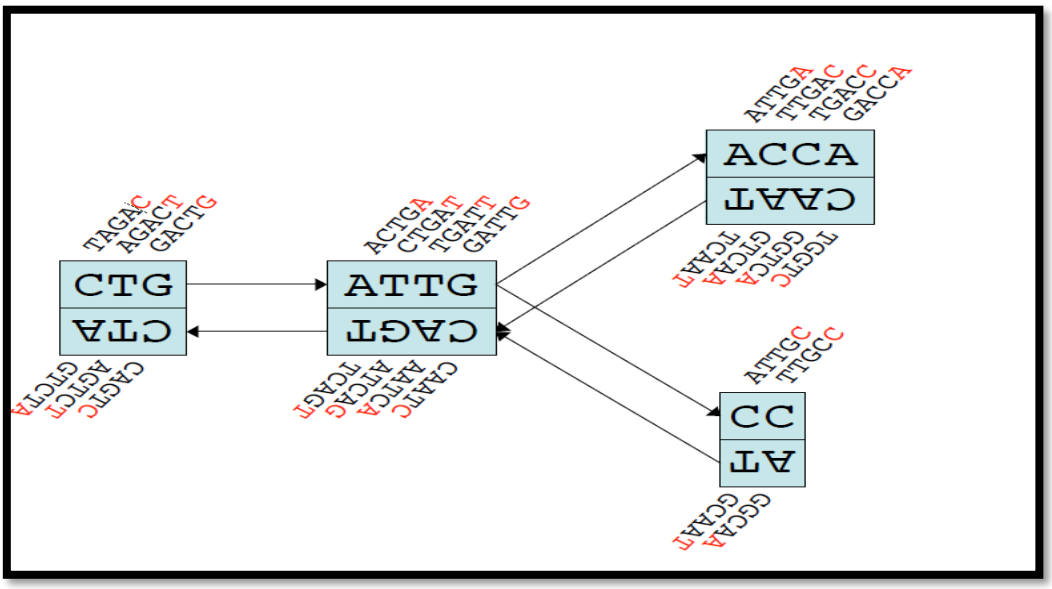
दोहरे छोर वाले शॉर्ट–रीड सीक्वेंसिंग प्रौद्योगिकियों के लिए असेम्बली

हाल ही में विकसित पायरोसीक्वेंसिंग – जैसी तकनीकें अत्यधिक आशाजनक हैं। तथापि, इनमें परिणामस्वरूप प्राप्त रीड्स की लंबाई वर्तमान सीक्वेंसिंग मशीनों द्वारा उत्पन्न रीड्स की तुलना में अत्यधिक कम होती है। सीक्वेंस रिपीट की लंबाई रीड्स की उपयोगिता को सीमित कर देती है। क्योंकि जब किसी भी सीक्वेंस रिपीट की लंबाई बढ़ जाती है तो इसे गैर समाधान के रूप में परिभाषित किया जाता है। विशेष रूप से छोटे रीड्स के संकलन में सबसे छोटा सामान्य सुपरस्ट्रिंग लक्ष्य के अत्यधिक सम्पीड़ित आकार का प्रतिनिधित्व करता है। रिपीटों की इस समस्या अर्थात् भिन्नतापूर्ण इंसर्ट लंबाई,

को हल करने के लिए दोहरे छोर रीड वाले प्रोटोकाल प्रस्तावित किए गए। खण्ड–बहु लक्षित क्लोन तथा लंबाई a+b:, या (समतुल्य बताए गए) d से d+w लंबाई के सभी खण्डों को अलग करने के लिए जैल इलेक्ट्रोफोरेसिस का उपयोग होता है। उपरोक्त समीकरण में इंटीजर d और w के मान निर्धारित होते हैं।

डी ब्रूजिन ग्राफ

वर्ष 1995 में, इड्यूरी और वाटरमैन ने असेम्बली को दर्शाने के लिए ग्राफ का उपयोग करना आरंभ किया। उन्होंने एक वैकल्पिक सीक्वेंसिंग तकनीक के लिए असेम्बली एल्गोरिथ्म प्रस्तुत किया जो संकरीकरण द्वारा सीक्वेंस को दर्शाता था जहां ओलिगोऐरे को k न्यूक्लियोटाइड संबंधी सभी शब्दों का पता लगाने के लिए इस्तेमाल किया जाता था और इन्हें ।. मर्स भी कहा जाता था जो किसी दिए गए जीनोम में उपस्थित थे। उनकी इस विधि में प्रत्येक पहचाने गए शब्द के लिए एक नोड सृजित करना और उसके बाद संबंधित नोडों को ओवरलैपिंग सम्बद्ध k मर्स से जोड़ना था। इसके बाद वे ओवरलैपिंग k- मर्स की श्रृंखला को रिपोर्ट कर सके जिससे सुस्पष्ट कोंटिंग उत्पन्न हुए क्योंकि इसमें शाखित कनेक्शन नहीं थे। यह सीक्वेंस ग्राफ डी ब्रूजिन ग्राफ कहलाता है जिसके द्वारा k–मर्स को कगारों के रूप में दर्शाया जाता है तथा ओवरलैपिंग ।.मर्स अपने छोरों से जुड़े हुए होते हैं। इसमें ग्राफ निर्माण, त्रुटि को दूर करने, मिश्रित लंबाई की असेम्बली व युग्मित–छोर की असेम्बली के लिए नए एल्गोरिथ्म होते हैं। तथापि, यह कार्यक्रम पुष्टता तथा आसानी से चलाने के लिए डिजाइन किया गया था। इसमें कुछ विशेष पहलू हैं : पहला, यह k- मर्स को कगारो की बजाय नोड्स पर मानचित्रित करता है। दूसरा, यह विलोम सम्पूरक सीक्वेंस को बाइ–ग्राफ (या–द्विदिशा वाले ग्राफ) को प्राप्त करने के लिए क्रमबद्ध करता है अर्थात् दूसरे शब्दों में ऐसा ग्राफ है जहां एक कोर इसके किसी भी छोर पर नोड में स्वतंत्र रूप से प्रविष्टि करती है या बाहर निकलती है। प्रत्येक नोड छ, ओवरलैपिंग k-मर्स की श्रृंखला को दर्शाती है। पास के k-मर्स, k-1 न्यूक्लियोटाइडों द्वारा ओवरलैप होते हैं। k-मर्स द्वारा उपलब्ध कराई गई सीमांत सूचना इसके अंतिम न्यूक्लियोटाइड में होती है। इन अंतिम न्यूक्लियोटाइड का सीक्वेंस नोड या `(N) का सीक्वेंस कहलाता है। इसलिए नोड का सीक्वेंस सम्बद्ध k-मर्स को अपूर्ण रूप से दर्शाता है। दूसरे शब्दों में k-मर्स के दो अलग–अलग सैट समान सीक्वेंस से युक्त दो अलग–अलग नोडों द्वारा दर्शाए जा सकते हैं। समान सीक्वेंस होने के बावजूद ये दोनों नोड अलग रखे जाते हैं तथा रीड्स का मानचित्रण उनमें निहित ।.मर्स के अनुसार किया जाता है। प्रत्येक नोड छ जुड़वां नोड ~N से जुड़ा होता है जो विलोम पूरक k-मर्स की विलोम श्रृंखला को दर्शाता है। इससे यह सुनिश्चित होता है कि विपरीत लड़ियों से लिए गए रीड्स जो एक–दूसरे को ओवरलैप करते हैं, उन्हें प्रयोग में शामिल कर लिया गया है। यह ध्यान देना महत्वपूर्ण है कि नोड या इसके जुड़वां के साथ सम्बद्ध सीक्वेंस एक–दूसरे के विलोम रूप से पूरक हों, यह आवश्यक नहीं है। नोड N तथा इसके जोड़े का मेल ब्लॉक कहलाता है। इसके पश्चात नोड में होने वाला कोई भी परिवर्तन इसके जोड़े के लिए भी समान रूप से लागू होता है। ब्लॉकों को इम्पलीसिट बाई–ग्राफ के नोड के रूप में माना जा सकता है। नोडों को एक निर्देशित कोर या चाप द्वारा जोड़ा जा सकता है। ऐसे मामले में किसी चाप के मूल नोड का अंतिम k-मर प्रथम डेस्टिनेशन नोड को ओवरलैप करता है। ब्लॉकों में सममितीय होने के कारण यदि चाप A से B नोड तक जाता है तो सममितीय ~B से ~A तक जाती है। किसी एक चाप में सुधार का अर्थ है कि इसके युग्म चाप में भी सममिति उत्पन्न होगी।



चित्र 2: डी ब्रूजिन ग्राफ के कार्यान्वयन का रेखाचित्र

एकल चतुर्भुज द्वारा दर्शाया गया प्रत्येक नोड सीधे ऊपर या नीचे सूचीबद्ध किए गए ओवरलैपिंग k-मर्स का प्रतिनिधित्व करता है (इस मामले में, k=5) प्रत्येक k-.मर का अंतिम न्यूक्लियोटाइड लाल रंग से दर्शाया गया है। चतुर्भुजों में बड़े अक्षरों में कॉपी किया गया अंतिम न्यूक्लियोटाइडों का सीक्वेंस नोड का सीक्वेंस है। नोड के नीचे या ऊपर सीधे जुड़े हुए युग्म नोड विलोम प्रतिपूरक ।.मर्स के विलोम श्रृंखला का प्रतिनिधित्व करते हैं। चापों को नोडों के बीच तीरों द्वारा दर्शाया गया है। चाप मूल का अंतिम ।.मर अपने गंतव्य के प्रथम स्थान पर ओवरलैप करता है। प्रत्येक में सममितीय चाप होता है। बाएं ओर के दो नोडों को सूचना में बिना किसी क्षति के एक साथ मिलाया जा सकता है क्योंकि ये एक श्रृंखला का निर्माण करते हैं।

- डी ब्रुजिन ग्राफ में ग्राफ के आर–पार पथों पर एक के साथ एक क्रमों के मानचित्र होते हैं। पथ से न्यूक्लियोटाइड सीक्वेंस का निष्कर्षण बिल्कुल सीधा होता है जो प्रथम नोड के आरंभिक ।.मर के रूप में दिया जाता है तथा इसे पथ के सभी नोडों में क्रमों के रूप में व्यक्त किया जाता है। प्रत्येक रीड के लिए ठीक एक पथ विद्यमान होता है जो सीक्वेंस के ।.मर्स से सम्बद्ध नोडों के माध्यम से क्रमबद्ध ढंग से आगे जाता है।
- दो ओवरलैपिंग सीक्वेंस दो पथों द्वारा अभिव्यक्त होते हैं जो एक–दूसरे को ओवरलैप करते हैं। पथों का परस्पर काट सीक्वेंसों के बीच के ओवरलैप से सम्बद्ध होता है। दो पथ टोपोलॉजी का उपग्राफ बनाते हैं जो सीक्वेंसों के बीच एलाइनमेंट के प्रकार से सीधे–सीधे जुड़ा हुआ होता है। यदि एक सीक्वेंस दूसरे को काट रहा होता है तो पथ भी अन्य पथ का उप–पथ हो जाता है। जब और क्रमों को जोड़ा जाता है तो उपरोक्त गुण प्रमाणित रहते हैं। इसका अर्थ है कि वे सभी क्रम जो समान सब्स्ट्रिंक की भागीदारी करते हैं वे रिपीटों के माध्यम से ओवरलैपिंग रीडों के सैटों की खोज करने में उपयोगी सिद्ध होते हैं क्योंकि ये एक ही पथ अपनाते हैं।

डी ब्रुजिन ग्राफ का प्रथम परिणाम यह है कि इसमें अत्यधिक विभिन्न लंबाइयों वाले सीक्वेंस को भी समायोजित किया जा सकता है। यह विशेष रूप से तब उपयोगी है जब मिश्रित लंबाई की सीक्वेंसिंग की जाती है या तुलनात्मक जीनोमिक्स में भी यह विशेष रूप से उपयोगी है। छोटे रीडों, लंबे रीडों, पूर्व एसेम्बल किए गए कोंटिग्स या अंतिम जीनोमों के लिए कोई भी तदर्थ अनुमान नहीं लगाना होता है। इसके अतिरिक्त पथ और सीक्वेंसों के बीच एक संबंध होने के कारण ओवरलैपिंग सीक्वेंस अनिवार्य रूप से समान पथ का अनुसरण करते हैं। इससे रीड्स के ओवरलैपिंग सैटों की निरंतरता के लिए खोज

करना आसान हो जाता है।

जटिल जीनोमों को असेम्बल करने से संबंधित मुद्दे और इससे जुड़ी समस्याएं

जीनोम असेम्बली एक अत्यंत कठिन कम्प्यूटेशनल समस्या है जो रिड्स की अधिक संख्या तथा समान सीक्वेंसों के कारण जो रिपीट कहलाते हैं, और भी जटिल हो जाती है। यह रिपीट हजारों न्यूक्लियोटाइड लंबे हो सकते हैं और कुछ विभिन्न स्थानों पर हजारों में होते हैं, विशेष रूप से पौधों और पशुओं के बड़े जीनोमों में तो ऐसा होता ही है।

फसल जीनोमों के अनुक्रमण के मामले में एक चुनौती जीनोमों के आकार तथा विभिन्न सीक्वेंसिंग विधियों द्वारा उत्पन्न किए गए रीड्स की लंबाई में अत्यधिक अंतर है। द्वितीय पीढ़ी की सीक्वेंसिंग और आधुनिक सैंगर सीक्वेंसिंग विधि से उत्पन्न किए गए छोटे रीडों के बीच पैमाने में 10–500 X का अंतर होता है। अनुक्रमित जीव जैसे–जैसे आकार में बढ़ते हैं, वैसे–वैसे असेम्बली कार्यक्रमों की जटिलता बढ़ती जाती है तथा जीनोम परियोजनाओं में इन समस्याओं को हल करने के लिए अत्याधुनिक कार्यनीतियां अपनाने की आवश्यकता होती हैः

- टैराबाइट सीक्वेंसिंग डाटा जिन्हें कम्प्यूटिंग क्लस्टरों पर ही संसाधित कर सकते हैं;
- समरूप और लगभग समरूप क्रम (जो रिपीट्स कहलाते हैं) जो सबसे खराब अवस्था हो सकती है। एल्गोरिथ्म की समय व अंतराल की जटिलता बढ़ने के साथ–साथ चरघातांकी रूप से बढ़ते जाते हैं; और
- सीक्वेंसिंग उपकरणों से फ्रेगमेंट में होने वाली त्रुटि जो असेम्बली को परिबद्ध कर सकती है।

सारणीः विद्यमान नवीन एसेम्बलर्स की सूची

नाम	प्रकार	प्रौद्योगिकियां	लेखक	कब अद्यतन हुआ
BySS	(बड़ा) जीनोम	सोलेक्सा, सोलिड	सिम्प्सन, जे और साथी	2008 / 2011
ALLPATHS-LG	(बड़ा) जीनोम	सोलेक्सा, सोलिड	ग्नेरे, एस और साथी	2011
AMOS	जीनोम	सेंगर, 454	साल्जबर्ग, ए. और साथी	2002 / 2008
एरापन–एम	मध्यम जीनोम (जैसे ई. कोलाई)	सभी	साहली, एम. और शिबुया , टी.	2011–2012
एरापन–एस	छोटे जीनोम (विषाणु और जीवाणु)	सभी	साहली, एम. और शिबुया, टी.	2011–2012
सेलेरा डब्ल्यूजीए असेम्बलर / सीएबीओजी	(बड़ा) जीनोम	सैंगर, 454, सोलेक्सा	मायर्स, जी. और साथी; मिलर जी. और साथी	2004 / 2010

सीएलसी / जीनोमिक्स वर्कबेंच और सीएलसी असेम्बली सैल	जीनोम	सैंगर, 454, सोलेक्सा, सोलिड	सीएलसी बायो	2008 / 2010 / 2011
कॉर्टेक्स	जीनोम	सोलेक्सा, सोलिड	इकबाल, जैड और साथी	2011
डी.एन.ए. बेसर	जीनोम	सैंगर, 454	हैराकल, बायोसॉफ़्ट एसआरएल	2013
डी.एन.ए. ड्रैगन	जीनोम	इल्युमिना, सोलिड, पूर्ण जीनोमिक्स, 454, सैंगर	सीक्वेंटी एक्स	2011
डीएनएनैक्सस	जीनोम	इल्युमिना, सोलिड, पूर्ण जीनोमिक्स	डीएनएनैक्सस	2011
एडेना	जीनोम	इल्युमिना	डी. हर्नांडेज, पी. फ्रांकोइस, एल. फैरिनेली, एम. ओस्टेरॉस और जे. स्क्रैजेंल	2008 / 2013
इयूलर	जीनोम	सैंगर, 454 (सोलेक्सा)	पैवजेनेर, पी. और साथी	2001 / 2006
इयूलर–एस.आर.	जीनोम	454, सोलेक्सा	चेइसन, एमजे और साथी	2008
फोर्ज	(बड़ा) जीनोम, ईएसटी, मैटाजीनोमस	454, सोलेक्सा, सोलिड, सैंगर	प्लाट, डीएम, एवर्स, डी.	2010

संदर्भः

1. बैट्जोगलोउ, एस., जैफे, डी.बी., स्टेंले, के, बटलर, जे, ग्नेरे, एस, माउसेली, ई. बर्जर, बी, मेसिरोव, जे.पी. और साथी (जनवरी 2002). 'एराक्ने' : एक होल जीनोम शॉटगन असेम्बलर' जीनोम रिसर्च 12(1):177–89. डीओआईः10.11.1 / जीआर 208902.पीएमसी 15525. पीएमआईडी 11779843.
2. बोइसवर्ट, सेबास्टियन, लेवियोलेटे, फ्रांकोइस, कोरबेइल, जैक्स (2010). 'रेः साइमलटेनियस एसेम्बली ऑफ रीड्स फ्राम , मिक्स ऑफ हाइ थ्रू पुट सीक्वेंसिंग टैक्नोलॉजीस'. जर्नल ऑफ कम्प्यूटेशनल

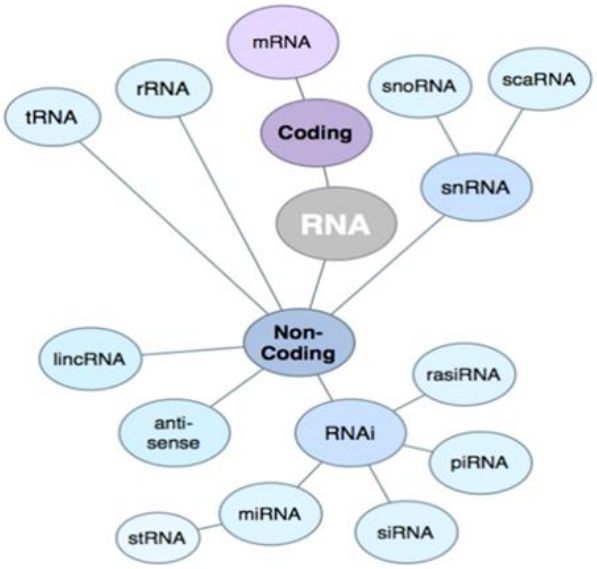
बायोलॉजी. 17(11) : 1519–3. डीओआईरु10.1089/सीएमबी.2009. 0238.पीएमसी 3119603. पीएमआईडी 20958248.
3. दोहम, जे.सी., लोटाज. सी.; बोरोडीना, टीत्र हिमेलबाउएर, एच. (नवम्बर 2007). 'सीएचएआरसीजीएस, ए फास्ट एंड हाइली एक्यूरेट शॉर्ट रीड असेम्बली एल्गोरिथ्म फॉर डीनोवो जीनोमिक सीक्वेंसिंग'. जीनोम रिसर्च 17(11): 1697–706. डीओआईः 10.1101/जीआर.6435207. पीएमसी 2045152. पीएमआईडी 17908823.
4. ह्यूस. एस.एम., ह्यूबर, जे.ए., मॉरीसन, एच.जी.,सोगिन, एम.एल. और वैल्च (डीएम) (2007). एक्यूरेसी एंड क्वालिटी ऑफ मैसिवली पैरलल डी.एन.ए. पाइरोसीक्वेंसिंग, जीनोम बायोल 8, आर 143.
5. मार्डिस, ई.आर. (2008). द इम्पेक्ट ऑफ नेक्स्ट जेनरेशन सीक्वेंसिंग टैक्नोलॉजी ऑन जेनेटिक्स, ट्रैंड्स जेनेट 24, 133–141.
6. माइकल सी. स्कार्टज, जान विटकोवस्की और डब्ल्यू रिचर्ड मैक कौम्बी (2012). करेंट चैलेंजिस इन डी नोवो प्लांट जीनोम सीक्वेंसिंग एंड असेम्बली. जीनोम बायोलॉजी, 13: 243.
7. मायर्स, ई. डब्ल्यू., सुटन, जीजी, डैल्चर, एएल, ड्यू, आई.एम., फासुले, डीपी, फलैनिगन, एमजे, क्राविड्स, एस.ए., मोवेरी, सीएम और साथी (मार्च 2000). 'ए होल जीनोम असेम्बली ऑफ ड्रोसोफिला' साइंस 287 (5461) : 2196–204. डीओआईः 10.01126/साइंस. 287.5461.2196. पीएमआईडी 1073113.
8. पॉप, एम. (2004) शॉटगन सीक्वेंस असेम्बली, एडवांस कम्प्यूटेशन 60, 193–248.7.
9. पॉप, एम. और स्लाजबर्ग, एस.एल. (2008). बायोइन्फोर्मेटिक्स चैलेंजिस ऑफ न्यू सीक्वेंसिंग टैक्नोलॉजी, ट्रैंड्स जेनेट 24, 142–149.
10. रोनागी, एम. उहलेन, एम और नाइरेन, पी. (1998). ए सीक्वेंसिंग मैथड बेस्ड ऑन रियल टाइम पाइरोफास्फेट, साइंस 281, 363–365.
11. झांग, डब्ल्यू, चैन जे., वांग वाई, तांग वाई, शांग जे, और साथी (2011). प्रैक्टिकल कैम्पेरीजन ऑफ डीनोवो जीनोम असेम्बली सॉफ्टवेयर टूल्स फॉर नैक्स्ट जेनेरेशन सीक्वेंसिंग टैक्नोलॉजिस. PLoS ONE 6(3):el7915.doi:10.1371/journal.pone.0017915.

डाटा एनालिसिस आर॰एन.ए- सीक्वैन्स

डॉ. सुधीर श्रीवास्तव
भा.कृ.अ.प.–भा.कृ.सां.अनु. संस्थान, नई दिल्ली–12

भूमिका

आगामी-पीढ़ी अनुक्रमण/ नेक्स्ट-जेनेरेशन सीक्वेंसिंग (एन.जी.एस.) तकनीक के आगमन ने जीनोमिक अध्ययन को बदल दिया है। एन.जी.एस. तकनीक का एक महत्वपूर्ण अनुप्रयोग ट्रांसक्रिप्टॉम का अध्ययन है। कोशिका में सभी आर.एन.ए. (RNA) अणुओं के पूर्ण संग्रह को ट्रांसक्रिप्टॉम कहा जाता है। विभिन्न प्रकार के आर.एन.ए. जिन्हें अब तक वर्गीकृत किया गया है, उन्हें चित्र 1 में दिखाया गया है। इन सभी अणुओं को ट्रांसक्रिप्टॉम कहा जाता है क्योंकि वे ट्रांसक्रिप्सन की प्रक्रिया द्वारा निर्मित होते हैं।



चित्र 1. विभिन्न प्रकार के आर.एन.ए.

एम.आर.एन.ए (mRNA) की एन.जी.एस. अर्थात् आर.एन.ए.-सेक (RNA-Seq) जैविक प्रयोगों में जीन अभिव्यक्ति (एक्सप्रेसन) को मापने के लिए एक मानक बन गया है। कम्प्यूटेशनल और सांख्यिकीय दृष्टिकोण से आर.एन.ए.-सेक डेटा (आंकड़ा) विश्लेषण सबसे संभावित शोध क्षेत्र रहा है जो ट्रांसक्रिप्टोमिक स्तर पर जीन की भूमिकाओं में एक अंतर्दृष्टि प्रदान कर सकता है। आर.एन.ए.-सेक डेटा उत्पन्न करने के लिए कई मशीनें/ प्रोटोकॉल उपलब्ध हैं, जैसे, इलुमिना (मिसेक, नेक्स्टसेक, हायसेक, नोवासेक), आयन टोरेंट (प्रोटॉन, पर्सनल जीनोम मशीन), सोलिड, रोच 454, आदि। आर.एन.ए.-सेक आंकड़ों का विश्लेषण माइक्रोएरे डेटा विश्लेषण से विभिन्न पहलुओं

में भिन्न होता है जैसे डेटा की प्रकृति, सामान्यीकरण के तरीके और डिफ्रेंसियल एक्सप्रेसन विश्लेषण।

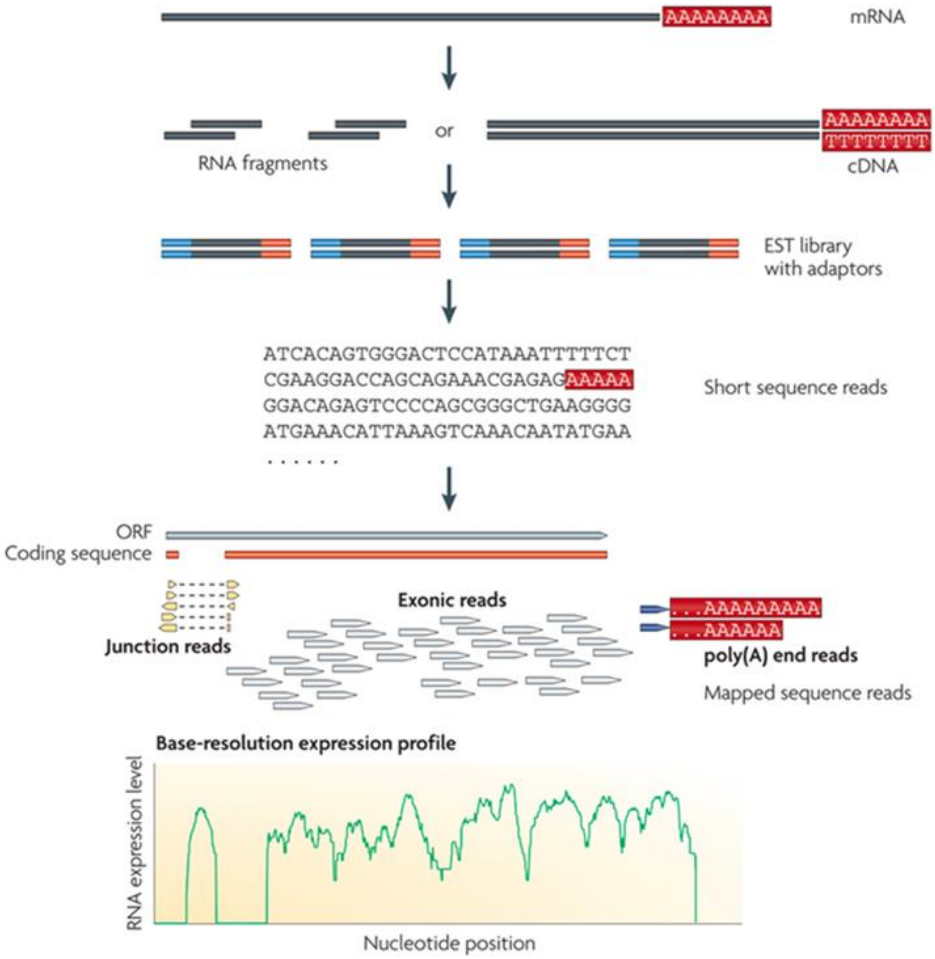
आर.एन.ए.-सेक के कई अनुप्रयोग हैं: ट्रांसक्रिप्टॉम/ आर.एन.ए. एक्सप्रेसन के स्तरों का परिमाणन, नई जीन की खोज, जीन एनोटेशन, विभिन्न स्थितियों के बीच डिफ्रेंसियली एबंडेंट/ एक्सप्रेस्ड फीचर्स (जीन्स/ ट्रांसक्रिप्ट्स/ एक्सॉन्स) का पता लगाना, स्प्लाइसिंग घटनाओं का पता लगाना, इंट्रॉन्स और एक्सॉन्स की सीमाओं की पहचान, इत्यादि।

आर.एन.ए.-सेक प्रयोग

आर.एन.ए.-सेक प्रयोग में कई महत्वपूर्ण चरण हैं: 1. डेटा उत्पादन (प्रयोगात्मक डिजाइन, सैम्पल संग्रह, सीक्वेंसिंग डिजाइन और गुणवत्ता नियंत्रण), 2. एक्सप्रेसन वेल्युस प्राप्त करने के लिए रीड्स (मैपिंग या एलाइनमेंट) का परिमाणन, 3. सामान्यीकरण; 4. डिफ्रेंसियल एक्सप्रेसन विश्लेषण। एक आम आर.एन.ए.-सेक प्रयोग को सारांशित करने के लिए मूल चरण निम्नानुसार हैं (कृपया चित्र 2 देखें):

- पहले शुद्ध आर.एन.ए. को सी.डी.एन.ए.(cDNA) में बदल दिया जाता है। फिर सीक्वेंसिंग लाइब्रेरी तैयार किया जाता है और एक एन.जी.एस. प्लेटफॉर्म पर सीक्वेंसिंग किया जाता है।
- सी.डी.एन.ए. अंशों के एक छोर (सिंगल-इंड) या दोनों छोर (पेयर्ड-इंड) से लाखों लघु सीक्वेंसिंग रीड्स उत्पन्न होते हैं।
- इन सीक्वेंस की मैपिंग संदर्भ (रिफरेंस) जीनोम से की जाती है।
- जाने हुए (ज्ञात) फीचर्स के लिए मैप की गई रीड की संख्या (रीड काउंट्स) को एक तालिका में दर्ज और संक्षेपित किया जाता है।

फीचर्स या तो जीन, ट्रांसक्रिप्ट (या अल्टरनेटिव ट्रांसक्रिप्ट), एलील स्पेसिफिक एक्सप्रेसन या एक्सॉन लेवल एक्सप्रेसन पर हो सकती हैं। उदाहरण के लिए, यदि F फीचर्स और N सैम्पल हैं, तो रीड काउंट्स की एक तालिका गैर- निगेटिव पूर्णांक का $F \times N$ मैट्रिक्स है।



चित्र 2. सामान्य आर.एन.ए.-सेक प्रयोग

आर.एन.ए.-सेक डेटा विश्लेषण

आर.एन.ए.-सेक विश्लेषण के लिए उपलब्ध कुछ ओपन सोर्स सॉफ्टवेअर इस प्रकार हैं:

रौ रीड डेटा (FASTQ फाइल्स) की गुणवत्ता जांच

- फास्ट क्यू. सी. (FastQC), एन.जी.एस.क्यू.सी. (NGSQC)

डेटा प्रीप्रोसेसिंग

- फास्टएक्स टूलकिट (FASTX toolkit), शॉर्टरीड (ShortRead), ट्रिम्मोमैटिक (Trimmomatic), सैमटूल्स (Samtools)

शॉर्ट रीड्स एलाइनर्स (Short reads aligners)

- बोटाई (Bowtie), टॉपहैट (TOPHAT), बी.डब्ल्यू.ए. (BWA), नोवोएलाइन (Novoalign), स्टार (STAR), आदि

डी नोवो अस्सेम्ब्लेर्स (*de novo* assemblers)

- सोपडीनोवो-ट्रांस (SOAPdenovo-Trans), ट्रांस-अबिस (Trans-AbySS), ट्रिनिटी (Trinity), स्पेड्स (SPAdes)

फीचर परिमाणन

- रौ रीड काउंट डेटा: एच.टी.सेक-काउंट (htseq-count), फीचरकाउंट्स (featureCounts)
- एक्सप्रेसन वेल्युस की परिमाणन करने के अन्य तरीके: कफ़लिंक्स (Cufflinks), स्ट्रिंगटाई (Stringtie), आर.एस.ई.एम. (RSEM), सेलफ़िश (Sailfish)

एक्सप्रेसन अध्ययन

- कफ़लिंक्स पैकेज (Cufflinks package)
- आर पैकेजेस (R packages): डी.ई.सेक (DESeq), डी.ई.सेक2 (DESeq2), एज.आर (edgeR), आदि

विजुअलाइज़ेसन

- कमेआरबंड (CummeRbund), आई.जीण्वी. (IGV), बेडटूल्स (Bedtools), यू.सी.एस.सी. जीनोम ब्राउज़र (UCSC Genome Browser), आदि

आर.एन.ए.-सेक रीड काउंट डेटा का एक उदाहरण

एक विशिष्ट आर.एन.ए.-सेक प्रयोग में, सैम्पल्स की सीक्वेंसिंग की जाती है और रीड्स की मैपिंग रिफरेंस जीनोम से की जाती है। प्रत्येक रिफरेंस जीन्स से मैप किए गए रीड्स की संख्या (रीड काउंट्स) की गणना की जाती है। मान लीजिये एक RNA-Seq प्रयोग में N सैम्पल्स हैं। आगे मान लीजिये कि स्थिति/ समूह C_i $(i = 1, 2)$ में j^{th} सैम्पल $(j = 1, 2, \dots, n_i)$ के जीन G_k $(k = 1, 2, \dots, K)$ में मैप किए गए रीड्स की संख्या Y_{ijk} है। आमतौर पर काउंट डेटा (Y_{ijk}) की मोडलिंग प्वाइजन डिस्ट्रिब्युसन या निगेटिव बायनोमियल डिस्ट्रिब्युसन द्वारा किया जाता है। एक काल्पनिक केस-कंट्रोल अध्ययन के लिए रीड काउंट की एक तालिका नीचे दी गई है (चित्र

3)।

	Conditions/ Treatment groups											
	C_1(Case)						C_2(Control)					
Samples → / Genes ↓	$S_{1,1}$	$S_{1,2}$	...	$S_{1,j}$	...	S_{1,n_1}	$S_{2,1}$	$S_{2,2}$	...	$S_{2,j}$	...	S_{2,n_2}
G_1	21	30	...	25	...	5	65	61	...	52	...	25
G_2	0	3	...	1	...	0	7	2	...	0	...	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_k	198	122	...	162	...	51	302	245	...	102	...	29
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_K	2	1	...	0	...	1	1	0	...	0	...	1

चित्र 3. एक काल्पनिक केस-कंट्रोल अध्ययन के लिए रीड काउंट की तालिका

डिफ्रेंसियल एक्सप्रेसन विश्लेषण के लिए विभिन्न आर पैकेजेस (R packages) उपलब्ध हैं जैसे कि एज.आर (edgeR), डी.ई.सेक (DESeq), डी.ई.सेक2 (DESeq2), आदि। डिफ्रेंसियल एक्सप्रेसन विश्लेषण करने से पहले सामान्यीकरण की आवश्यकता होती है। सामान्यीकरण के अनेक तरीके हैं जैसे कि आर.पी.के.एम. (रीड्स अलाइन्ड पर किलोबेस ऑफ एक्सॉन पर मिलियन रीड्स मैप्ड), एफ.पी.के.एम. (फ्रेगमेंट्स अलाइन्ड पर किलोबेस ऑफ एक्सॉन पर मिलियन फ्रेगमेंट्स मैप्ड) और टी.पी.एम. (ट्रांसक्रिप्ट्स पर किलोबेस मिलियन)। ये तरीके डेटा के सामान्यीकरण के लिए जीन की लंबाई और सीक्वेंसिंग डेप्थ का उपयोग करते हैं।

निष्कर्ष

आर.एन.ए.-सेक अभी भी उपयोग में है और पहले से विकसित ट्रांसक्रिप्टॉमिक विधियों पर इसके लाभ स्पष्ट हैं। मगर आर.एन.ए.-सेक प्रयोगों के उपयोग से जुड़ी अनेक चुनौतियां हैं जैसे लाइब्रेरी का निर्माण, जैव सूचना विज्ञान की समस्या (बड़े डेटा सेट का संचयन, पुनर्प्राप्ति और प्रोसेसिंग; मैपिंग और असेंबली की समस्या), सीक्वेंस/ ट्रांसक्रिप्टॉम कवरेज बनाम लागत, ट्रांसक्रिप्टॉमिक विश्लेषण (इंट्रॉन्स और एक्सॉन्स की सीमाओं की पहचान के साथ-साथ नई जीन की खोज के लिए जीन मैपिंग; स्प्लाइसिंग घटनाओं का पता लगाना; जटिल प्रयोगों में जीन एक्सप्रेसन का अध्ययन करने के लिए ट्रांसक्रिपटॉम/ आर.एन.ए. एक्सप्रेसन के स्तरों का परिमाणन), इत्यादि। आने वाले भविष्य में यह मौजूदा तकनीक पर अधिक सुधार के साथ और बेहतर हो जाएगा तथा अन्य अनुप्रयोगों के

लिए यह माइक्रोएरे जैसी तकनीक की जगह ले लेगा।

संदर्भ

1. वांग जेड., गेरस्टीन एम., स्नाइडर एम. (2009). आर.एन.ए.-सेक: ट्रांसक्रिप्टॉमिक्स के लिए एक क्रांतिकारी टूल, *नेचर रीव्यू जेनेटिक्स*, **10 (1)**, 57-63।
2. एंडर्स एस., ह्यूबर डब्ल्यू. (2010). सीक्वेंस काउंट डेटा के लिए डिफ्रेंसियल एक्सप्रेसन विश्लेषण,*जीनोम बायोलोजी*, **11**, R106।
3. मोर्टाज़ावी ए., विलियम्स बी.ए., मैक्यू के., शेफ़र एल., और वोल्ड बी. (2008). आर.एन.ए.-सेक द्वारा मैमलियन ट्रांसक्रिप्टॉम्स का मैपिंग और परिमाणित करना, *नेचर मेथड्स*, **5 (7)**, 621-628।
4. शेंड्योर जे., जी एच. (2008). नेक्स्ट-जेनेरेशन आर.एन.ए. सीक्वेंसिंग, *नेचर बायोटेक्नोलॉजी*, **26**, 2514-2521।
5. ब्रायन जे. एच. और माइकल सी. जेड. (2010). आर.एन.ए.-सेक विश्लेषण को आगे बढ़ाना, *नेचर बायोटेक्नोलॉजी*, **28**, 421-423।

गफज़ी रैखिक समाश्रयण तथा इसके अनुप्रयोग

हिमाद्रि घोष

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली–110012

1. **प्रस्तावनाः** यह पूर्ण रुप से प्रमाणित है कि कृषि विज्ञान भौतिक एवं रसायन विज्ञान जोकि कठिन विज्ञान है, से भिन्न एक सरल विज्ञान है। कृषि विज्ञान में अन्तर्निहित विषय और/या व्याख्यात्मक चर और/या प्रतिक्रियात्मक चर में "अपरिषुद्धता" या "अनिष्चिता" या "अस्पष्टता" की कुछ मात्रा हमेषा रहती है। इसलिये अत्यधिक यथार्थवादी मॉडलिंग के लिये इस पहलू को परम्परागत मॉडलों जैसे बहुरैखिक समाश्रयण मॉडल की श्रेणी में समाविष्ट करने की आवष्यकता है।

2. **फज़ी रैखिक समाश्रयण कार्यपद्धतिः** परम्परागत समाश्रयण विष्लेषण में, अवलोकित एवं अनुमानित मानों के बीच यादृच्छिक त्रुटियों के कारण विचलन की कल्पना की जाती है। परन्तु, ये अक्सर पद्धति की अनियमित संरचना या अवलोकन के कारण होता है। अतः इस प्रकार के समाश्रयण मॉडल में अनिष्चितता "अस्पष्टता"श्बन जाती है न कि यादृच्छिकता। फज़ी रैखिक समाश्रयण अध्ययन विस्तार से दो कार्यपद्धतियों में वर्गीकृत किया जा सकता है, अर्थात् ;पद्ध रैखिक प्रोग्रामिंग ;स्च्द्ध – पर आधारित विधि, और ;पपद्ध फज़ी लीस्ट स्क्वेर ;थ्स्ैद्ध विधियाँ। प्रथम कार्यपद्धति में, जो कि वर्ष 1982 में टनाका ;ज्दांद्ध एवं अन्य द्वारा प्रस्तावित की गयी थी, फज़ी रैखिक मॉडल के प्राचल

$$y = A_0 + A_1 X_1 + ApXp \qquad ;1द्ध$$

जहाँ

$$Ai = \left(a_{ic}, a_{jw}\right), y = \left(y_C, y_W\right)$$

हैं, मॉडल आँकड़ों को संगठित कर "कुल अस्पष्टता" को न्यूनतम कर इस आधार पर अनुमानित किये गए कि सभी आँकड़े प्रतिक्रियात्मक चरों के दायरे में सीमित है। इसे स्च समस्या के रुप मे देखा जा सकता है और इसे श्सिम्पलेक्स प्रक्रियाश् की मदद से हल किया जा सकता है। कई सॉफ्टवेयर पैकेज़ जैसे एस.ए.एस ;।)ए एल.पी.88 (स्च्88) और एल.आई.एन.डी.ओ. (स्प्छक्त्)ए इस समस्या को हल करने के लिये उपलब्ध हैं। किसी भी मानक स्प्रेड शीट पैकेज़, जैसे माइक्रोसॉफ्ट एक्सेल के द्वारा भी एल.पी. समस्या को हल किया जा सकता है।

वर्ष (2003) में काण्डाला तथा प्रज्ञेषु ने उपर्युक्त कार्यप्रणाली की उपयुक्तता को प्रमाणित किया, जब दो व्याख्यात्मक चर राषियां (अर्थात् पौधे की ऊंचाई तथा पत्ती का क्षेत्रफल सूचंकाक) तथा प्रतिक्रियात्मक चर राषि (शुष्क द्रव्य संचय) सभी स्पष्ट है लेकिन तथ्य में अन्तर्निहित विषय कें अस्पष्ट होने की कल्पना की जाती है। यह देखा गया कि फज़ी रैखिक समाश्रयण मॉडल में मानित अन्तराल की चौड़ाई बहुरैखिक समाश्रयण मॉडल के लिये ली गयी चौड़ाई की तुलना में काफी कम थी। पहले वर्ष (2002) में काण्डाला तथा प्रज्ञेषु ने इसी प्रकार के परिणाम प्राप्त किये थे जब दो व्याख्यात्मक चर राषियां, तथा सामान्यकरण अन्तर वनस्पति सूचकांक (एन.डी.वी.आई.) तथा वनस्पति सूचकांक अनुपात (आर.वी.आई.) उच्चीय सहसंबंध की स्थिति में हैं।

इसके अलावा, मछलियों की प्रजाति में आयु–लम्बाई सम्बन्ध की गणना करने के लिये, प्रतिक्रियात्मक चर राषि (लम्बाई) सामान्यतः समान आयु की विभिन्न मछलियों के अन्तराल में पायी जाती है। वर्ष 2004

में काण्डाला तथा प्रज्ञेषु ने घोंघा मछली में आयु लम्बाई सम्बन्ध की गणना करने के लिये फज़ी वॉन बर्टरलैनफी ;अवद ठमतजंसंदलिद्ध की ग्रोथ मॉडल का अवलोकन एफ.एल.आर. कार्यपद्धति द्वारा किया। यह देखा गया कि पारम्परिक सांख्यिकीय विधियाँ उस स्थिति का प्रतिपादन करने में सक्षम नहीं हैं, जिसमें प्रतिक्रियात्मक चर राषि अन्तराल में है। प्रतिक्रियात्मक चर राषि के अन्तराल मूल्यों को स्पष्ट बनाने का तरीका यह है कि उनका माध्य या मोड लिया जाये जिसमें प्रसार की महत्वपूर्ण जानकारी लुप्त हो जाने की संभावना है।

परन्तु, टनाका के प्रस्ताव की आलोचना यह है कि यह ठोस सांख्यिकीय सिद्धांतों पर आधारित नहीं है। दूसरी खामी चांग और अय्यूब (2001) ने तथा डी' उसरो (2003) ने यह निकाली कि जैसे–जैसे आँकड़ों की संख्या बढ़ती है, एल.पी. में अवरोधों की संख्या भी उसी अनुपात में बढ़ती है, जिसके कारण परिणाम निकालने में संगणात्मक कठिनाइयाँ आती हैं।

द्वितीय प्रस्ताव (1988) में डायमंड द्वारा खोजी गई फज़ी लीस्ट वर्ग (एफ.एल. एस.) विधि पर आधारित है जिसके नाम से ही स्पष्ट होता है कि लीस्ट स्क्वेर पद्धति का फज़ी विस्तार फज़ी संख्या की दूरी के नये परिभाषित अन्तराल पर आधारित है। काण्डाला तथा प्रज्ञेषु (2004) ने कुछ मछलियों की प्रजाति की लम्बाई–भार आँकड़ों का अनुमान लगाने के लिये प्रचलित "एलैमिट्रिक मॉडल" के लिये इस कार्य पद्धति का प्रयोग का प्रयोग किया। यद्यपि इस विधि की एक कमी यह है कि जैसे–जैसे व्याख्यात्मक चर राषि का परिमाण बढ़ता है, अनुमानित प्रतिक्रिया का फैलाव बढ़ता है जबकि अवलोकित प्रतिक्रिया का फैलाव स्थिर होता है या घटता है। इस स्थिति से उभरने के लिये काओ और च्यु (2002) ने एफ.एल.एस. पद्धति द्वारा एफ.एल.आर. का अवलोकन करने के लिये श्टू–स्टैजश पद्धति का प्रस्ताव रखा, यह पद्धति डायमंड पद्धति से श्रेष्ठ सिद्ध हुई। हाल ही में सिंह एवं अन्य (2007) ने इस पद्धति पर पूर्ण रूप से विचार विमर्ष किया तथा इसके अनुप्रयोग के लिये, उपयुक्त संगणक प्रोग्राम श्अरैखिक प्रोग्रामिंग सौलवर एल.आई.एन.जी.ओ, संस्करण 8श सॉफ्टवेयर पैकेज़ में विकसित किये गए। विषेष रूप से फसल पैदावार फज़ी आकलन के लिये श्फज़ी लीस्ट सक्वेयरश पद्धति द्वारा प्राचलों का आकलन किया गया। जैसे उदाहरण के लिए हरियाणा, भिवानी जिले के किसानों के आकलन के आधार पर ब्लाक स्तर के आकलनों के क्रम में ज्वार फसल की पैदावार के आँकड़ों पर कार्यप्रणाली का प्रयोग किया गया । कार्य के मूल्यांकन जांच का प्रयोग करके योग्यता स्तर के अनुकूलतम मान पर संभावना, आवष्यकता और न्यूनतम परिणामों की तुलना की गई। अन्ततः व्याख्यात्मक चर के विभिन्न स्पष्ट मूल्यों के लिये तथा उनके समरुप अन्तराल में दिये गये प्रतिक्रियात्मक चर के लिये घोंघा मछली के आयु–लम्बाई आँकड़ों का अवलोकन वॉन बर्टरलैनफी ;अवद ठमतजंसंदलिद्ध ग्रोथ मॉडल द्वारा किया गया। ज्वार आयस्टर आयु लम्बाई आँकड़ों के लिये पूर्ण किया गया। यंग तथा लीयू (2003) ने एक एफ.एल. आर. मॉडल में आउटलायर की उपस्थिति के विरूद्ध सुदृढ़ कलन विधि विकसित की। परन्तु, इस कार्यपद्धति को अभी कृषि के क्षेत्र के आँकड़ों पर प्रयोग करने की आवष्यकता है।

सन्दर्भ

बकले,जे.जे. तथा फ्यूरिंग,टी. (2000). लीनियर तथा नॉन–लीनियर फज़ी रिग्रेषनः रिवोल्यूषनरी एल्गोरिथ्म साल्यूषन. फज़.सैट्स सिस्ट.,**112**, 381–94

चांग,वाई,ओ. तथा अय्यूब,बी.एम. (2001). फज़ी रिग्रेषन मैथड्स–ए कम्पैरेटिव असेसमेन्ट. फज़.सैट्स सिस्ट.,**119**, 187–203

डायमन्ड,पी. (1988). फज़ी लीस्ट स्क्वायर्स,इन्फोर्म, साई.,**46**, 141–57

डी'उसरो,पी. (2003). लीनियर रिग्रेषन एनालाइसिस फॉर फज़ी/क्रिस्प इनपुट एण्ड फज़ी/क्रिस्प आउटपुट डाटा.काम्प.स्टेटिस्ट.डैट.एनल.,**42**, 47–72

होगं,डी.एच. एण्ड ह्ववांग,सी.. (2003). स्पोर्ट वेक्टर फज़ी रिग्रेषन मषीन्स फज़.सैट्स सिस्ट.,**138**, 271–81

कान्डाला,वी.एम. तथा प्रज्ञेषु (2002). फज़ी रिग्रेषन मैथड्लोजी फॉर क्रौप यील्ड फोरकास्टिंग यूजिंग रिमोटली सेन्सड डाटा.जे.इन्ड.सोस.रेम.सेन्सिंग,**30**, 191–95

कान्डाला,वी.एम. एण्ड प्रज्ञेषु (2003). एप्पलीकेषन ऑफ फज़ी रिग्रेषन मैथड्लोजी इन एग्रीकल्चर. इन्ड. जे.एग्रिक.साई.,**73,** 456–58

कान्डाला,वी.एम. एण्ड प्रज्ञेषु (2004). फज़ी वॉन बर्टरलैनफी ग्रोथ मॉडल फॉर डेटरमानिंग ऐज–लेन्थ रिलेशलनशिप. इन्ड.जे.फिश.,**51,** 55–59

कान्डाला,वी.एम. एण्ड प्रज्ञेषु (2004). फिटिंग एलोमिट्रिक मॉडल यूजिंग फज़ी लीस्ट स्क्वायर्स. जे.मार. बायोल.एएसएसन.,**46,** 120–23

काओ,सी. एण्ड च्यू,सी.एल. (2002). ए फज़ी लीनियर रिग्रेषन मॉडल विद बैटर एक्सप्लेनेटरी पावर.फज़. सैट्स.सिस्ट.,**126,** 401–09

सिंह,आर.के., घोष,एच., एण्ड प्रज्ञेषु (2008). पॉसिबिलिटी तथा नेसेसिटी मेज़र्स फॉर फज़ी लीनियर रिग्रेषन एनालाइसिसः एन एप्पलीकेषन.जे. इन्ड.सोस.एजी.स्टेट.,**62,** 19–25

सिंह,आर.के., प्रज्ञेषु, एण्ड घोष,एच. (2008). ए टू–स्टेज फज़ी लीस्ट स्क्वायर्स प्रासीजर्स फॉर फिटिंग लीनियराइज्ड वॉन बर्टरलैनफी ग्रोथ मॉडल. इन्ड.जे.फिष. **55,** 235–40

टनाका,एच., यूजीमा,एस., एण्ड एसाया,के. (1982). लीनियर रिग्रेषन एनालाइसिस विद फज़ी मॉडल. आई.ई.ई.ई. ट्रान्स. सिस्टमस मैन. साइबरनेट.,**12,** 903–07

यंग,एम.एस. एण्ड लीयू,एच.एच.(2003). फज़ी लीस्ट स्क्वायर्स एल्गोरिथ्म फॉर इन्टरएक्टिव फज़ी लीनियर रिग्रेषन मॉडलस. फज़.सैट्स.सिस्ट.,**135**, 305–16

फजी लीनियर रिग्रेशन पद्धति का उपयोग

1. निम्नलिखित डेटा वर्षा आधारित ग्रीनग्राम की उत्पादकता पर सल्फर युक्त उर्वरकों का प्रभाव देता है। प्रतिक्रिया चर (Y) शुष्क पदार्थ संचय है और व्याख्यात्मक चर हैं: संयंत्र ऊंचाई (X1) और पत्ती क्षेत्र सूचकांक (X2)। मॉडल को अंतर्निहित फजी मानते हुए एसएएस सॉफ्टवेयर पैकेज में उपलब्ध रैखिक पद्धति का उपयोग कर डेटा के लिए फजी लीनियर रिग्रेशन (एफएलआर) फिट कराए और लीनियर रिग्रेशन (एमएलआर) पर इसकी श्रेष्ठता दिखाए :

Y (g/m^2):	247.32	324.52	364.56	328.44	349.48	339.92	320.48	357.16
X_1 (cm):	60.41	61.08	64.98	64.16	62.99	65.20	63.24	67.19
X_2 :	3.74	4.80	5.71	5.27	5.45	5.34	5.11	5.66

parameters of FLR model

$$Y = A_0 + A_1X_1 + \ldots + A_pX_p$$

are ac1 ac2 ac3 and width aw1 aw2 aw3 .

SAS COMMANDS:

```
Title 'Method of least square';
Ods rtf file= 'resultls.rtf';
data plant;
input y x1 x2;
cards;
247.32  60.41  3.74
324.52  61.08  4.8
364.56  64.98  5.71
328.44  64.16  5.27
349.48  62.99  5.45
339.92  65.2   5.34
320.48  63.24  5.11
357.16  67.19  5.66
;
ods output ParameterEstimates=parms;
proc reg;
 model y=x1 x2;
output out=all;
proc print data=all;
 run;
/* To calculate lower and upper limits*/
 proc iml;
 use plant;
 read all into abc;
 a=abc[,2];
 b=abc[,3];
 use parms;
 read all into xyz;
 x=xyz[,2];
 y=xyz[,3];
 n=8;
 do i=1 to n;
lb=(x[1]-y[1])+((x[2]-y[2])*a[i])+((x[3]-y[3])*b[i]);
```

```
    ub=(x[1]+y[1])+((x[2]+y[2])*a[i])+((x[3]+y[3])*b[i]);
    w=ub-lb;
    lim=lb||ub||w;
          limit=limit//lim;
          width=width//w;
    end;

          average =mean(width);
    print limit;
          print average;
          quit;
          ods rtf close;
```

/*Method of Fuzzy regression (FR) (OPTMODEL)*/

```
        Title 'Linear programming';
data plant;
input  y  x1  x2;
datalines;
247.32  60.41    3.74
324.52  61.08    4.80
364.56  64.98    5.71
328.44  64.16    5.27
349.48  62.99    5.45
339.92  65.20    5.34
320.48  63.24    5.11
357.16  67.19    5.66
;
run;

proc optmodel;
  ods trace on;
  ods output PrintTable=parms;
  set j= 1..8;
  number y{j}, x1{j}, x2{j};
  read data plant into [_n_]  y  x1  x2;
   number n init 8; /* Total number of Observations*/
 /* Decision Variables*/
   var aw{1..3}>=0;     /*Theses three variables are bounded*/
   var ac{1..3};       /* These three variables are not bounded*/

 /* Objective function*/
   min z1=  aw[1] * n + sum{i in j} x1[i] * aw[2] + sum{i in j} x2[i] * aw[3];

  /*Linear Constraints*/
  con c{i in 1..n}:    ac[1]+x1[i]*ac[2]+x2[i]*ac[3]-aw[1]-x1[i]*aw[2]- x2[i]*aw[3] <= y[i];
con c1{i in 1..n}: ac[1]+x1[i]*ac[2]+x2[i]*ac[3]+aw[1]+x1[i]*aw[2]+x2[i]*aw[3] >= y[i];

    expand;    /* This provides all equations */
    solve;
    print ac  aw ;
          proc print data=parms;
          run;
          ods trace off;
```

```
quit;
proc iml;
        use plant;
        read all into abc;
        a=abc[,2];
        b=abc[,3];
        use parms;
        read all into xyz;
        x=xyz[,2];
        y=xyz[,3];
        do i=1 to 8;
  lb=(x[1]-y[1])+((x[2]-y[2])*a[i])+((x[3]-y[3])*b[i]);
  ub=(x[1]+y[1])+((x[2]+y[2])*a[i])+((x[3]+y[3])*b[i]);
  w=ub-lb;
        width=width//w;
  lim=lb||ub||w;
        limit= limit//lim;
        end;
        print limit;
        avg=width[:];
        print avg;
       quit;
```

RESULTS

Problem 1

Method of Multiple linear regression (MLR)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	186.10038	107.62844	1.73	0.1444
x1	1	-3.01638	2.16085	-1.40	0.2216
x2	1	65.21831	7.57036	8.61	0.0003

Output: Y=186.10 – 3.02 X1 + 65.22 X2
Standard Errors (107.63) (2.16) (7.57)

Method of Fuzzy regression (OPTMODEL)

ac1= 217.08 ac2= -3.0657 ac3= 59.734 aw1= 7.9678 aw2=0 aw3=0

Table1: Fitting of FLR by linear programming approach

Multiple Linear Regression (MLR) Model			Fuzzy Linear Regression (FLR) Model		
Lower limit	Upper limit	Width	Lower limit	Upper limit	Width
-18.68	514.27	532.96	247.32	263.25	15.93
38.96	590.76	551.90	308.58	324.52	15.93
71.22	653.76	582.53	350.99	366.92	15.93
50.10	622.43	572.33	327.22	343.15	15.93
66.54	636.54	569.99	341.56	357.49	15.93
48.76	626.64	577.88	328.21	344.14	15.93

45.64	611.57	565.93	320.48	336.41	15.93
56.90	648.23	591.33	341.22	357.16	15.93

औसत चौड़ाई **568.11** औसत चौड़ाई **15.93**

एफएलआर मॉडल की छोटी चौड़ाई एमएलआर मॉडल पर श्रेष्ठता दर्शाती है।