Lab exercises. Rare variant concepts and tools with Bioconductor.

VJ Carey

March 9, 2012

1 Acknowledgment

The material covered in this lab depends heavily on software designs and creations of Martin Morgan and Sean Davis; the *IRanges*, *GenomicRanges*, *BSgenome* and *AnnotationDbi* infrastructures are also critical for allowing these problems to be stated and solved concisely, so great thanks are due to Patrick Aboyoun, Marc Carlson, Mike Lawrence and Hervé Pages.

2 Resources

Using the 0.1.6 version of samtools, we created a pileup of 1000 genomes reads from NA19240's solexa image:

```
samtools pileup -cvf \
  human_b36_female.fa NA19240.chrom17.SLX.maq.SRP000032.2009_07.bam > \
  n240_17.pup
```

This text file is available to you for parsing as follows:

```
> library(ind1KG)
> library(Rsamtools)

> pup17 <- gzfile(system.file("pileups/n240_17.pup.gz", package="ind1KG"))
> c17p.i <- readPileup(pup17, variant="indel")
> levels(seqnames(c17p.i))

[1] "17"

> seqlevels(c17p.i) = gsub("17", "chr17", seqlevels(c17p.i))
> c17p.i
```

```
GRanges with 22640 ranges and 11 elementMetadata values:
          seqnames               ranges strand   | referenceBase consensusBase
             <Rle>            <IRanges>  <Rle>   |      <factor>      <factor>
      [1]    chr17       [55518, 55518]      *   |             A             A
      [2]    chr17       [55994, 55994]      *   |             T             T
      [3]    chr17       [56014, 56014]      *   |             A             R
      [4]    chr17       [57801, 57801]      *   |             G             G
      [5]    chr17       [59631, 59631]      *   |             C             C
      [6]    chr17       [62489, 62489]      *   |             G             G
      [7]    chr17       [62491, 62491]      *   |             C             C
      [8]    chr17       [62495, 62495]      *   |             C             C
      [9]    chr17       [62498, 62498]      *   |             A             A
      ...      ...                  ...    ... ...           ...           ...
  [22632]    chr17 [78614975, 78614975]      *   |             A             A
  [22633]    chr17 [78623969, 78623969]      *   |             T             T
  [22634]    chr17 [78623971, 78623971]      *   |             G             G
  [22635]    chr17 [78630685, 78630685]      *   |             C             C
  [22636]    chr17 [78632824, 78632824]      *   |             G             G
  [22637]    chr17 [78632825, 78632825]      *   |             A             A
  [22638]    chr17 [78632827, 78632827]      *   |             C             C
  [22639]    chr17 [78632832, 78632832]      *   |             C             C
  [22640]    chr17 [78654090, 78654090]      *   |             T             T
          consensusQuality snpQuality maxMappingQuality  coverage   alleleOne
                 <integer>  <integer>         <integer> <integer> <character>
      [1]              98          0                47        31          -T
      [2]              70          0                52        20          -A
      [3]              37         37                46        18          +G
      [4]              62          0                57        27          +A
      [5]              72          0                60        21         -AT
      [6]              61          0                56        36         +AC
      [7]             123          0                56        32         +AA
      [8]             107          0                56        33         +AG
      [9]             106          0                56        38         +CG
      ...             ...        ...               ...       ...         ...
  [22632]             108          0                54        27          +C
  [22633]             123          0                53        32           *
  [22634]             106          0                53        33           *
  [22635]              29          0                26        10           *
  [22636]              79          0                52        29          +A
  [22637]              89          0                54        29          +C
  [22638]              88          0                54        32           *
  [22639]              91          0                55        33           *
```

```
  [22640]               43          0                48        14          -C
          alleleOneSupport   alleleTwo alleleTwoSupport additionalIndels
                 <integer> <character>        <integer>        <integer>
      [1]                2           *               29                0
      [2]                6           *               14                0
      [3]                7           *               11                0
      [4]               17           *               10                0
      [5]               11           *               10                0
      [6]                7           *               29                0
      [7]                1           *               31                0
      [8]                1           *               32                0
      [9]                1           *               37                0
      ...              ...         ...              ...              ...
  [22632]                1           *               26                0
  [22633]               31          +G                1                0
  [22634]               24          +A                9                0
  [22635]                9          -G                1                0
  [22636]                1           *               28                0
  [22637]                2           *               27                0
  [22638]               31          +A                1                0
  [22639]               31          -A                2                0
  [22640]                3           *               11                0
  ---
  seqlengths:
   chr17
      NA
```

Information on dbSNP SNP locations is available in

```
> library(SNPlocs.Hsapiens.dbSNP.20090506)
> c6 <- getSNPlocs("chr6")
> head(c6, 5)

  RefSNP_id alleles_as_ambig   loc
1   6922869                Y 92596
2   6905277                S 92646
3  71545186                M 92724
4  71545187                M 92909
5   6923601                Y 92941
```

Resource-oriented exercise: After reviewing the 'Setup' material in the next section, estimate the frequencies of dbSNP-catalogued SNP per base pair in intronic vs. exonic DNA on chromosome 17. Estimate frequencies stratified by GC content (i.e., tabulate by 0, 1, 2 bases G or C in SNP).

Special design exercise (attempt only after all other exercises below have been solved correctly): Consider alternative representations of the SNP location/value data. The allele data could be represented as a single *DNAString*, and the location information as an *IRanges* instance. Assess the resource consumption and query resolution performance of these representations in comparison to the existing data.frame. Consider also a representation rooted in an RDBMS such as SQLite.

3 Exercises

3.1 Check for coding indels and SNP

- Setup: According to data distributed by Shendure, NA19240 has two copies of a triple insert in the coding region of CDRT4. Verify.

```
> library(org.Hs.eg.db)
> egid <- get("CDRT4", revmap(org.Hs.egSYMBOL))
> kgid <- get(egid, org.Hs.egUCSCKG)
> library(GenomicFeatures)
> library(TxDb.Hsapiens.UCSC.hg18.knownGene)
> txdb = TxDb.Hsapiens.UCSC.hg18.knownGene
> txloc <- transcripts(txdb)
> cdrt4txloc <- txloc[elementMetadata(txloc)$tx_name %in% kgid]
> subsetByOverlaps(c17p.i, cdrt4txloc)

GRanges with 12 ranges and 11 elementMetadata values:
       seqnames               ranges strand | referenceBase consensusBase
          <Rle>            <IRanges>  <Rle> |      <factor>      <factor>
   [1]    chr17 [15283313, 15283313]      * |             G             G
   [2]    chr17 [15283315, 15283315]      * |             T             T
   [3]    chr17 [15284249, 15284249]      * |             C             C
   [4]    chr17 [15286452, 15286452]      * |             G             G
   [5]    chr17 [15286466, 15286466]      * |             A             A
   [6]    chr17 [15288786, 15288786]      * |             G             G
   [7]    chr17 [15290129, 15290129]      * |             T             T
   [8]    chr17 [15295510, 15295510]      * |             C             C
   [9]    chr17 [15304438, 15304438]      * |             C             C
  [10]    chr17 [15304444, 15304444]      * |             A             A
  [11]    chr17 [15306197, 15306197]      * |             G             G
  [12]    chr17 [15306201, 15306201]      * |             G             G
       consensusQuality snpQuality maxMappingQuality  coverage   alleleOne
              <integer>  <integer>         <integer> <integer> <character>
   [1]               93          0                59        38          -T
```

```
  [2]              135          0                59        36          -A
  [3]               42          0                55        19        +CTT
  [4]                3          0                47        24          -A
  [5]               90          0                52        31           *
  [6]              115          0                54        40           *
  [7]               88          0                58        40          -C
  [8]               31          0                57        21   +CCATGGCT
  [9]               86          0                58        31        +TAA
 [10]               80          0                58        29        +TAC
 [11]               54          0                56         9          +T
 [12]               31          0                58         8          +C
      alleleOneSupport   alleleTwo alleleTwoSupport additionalIndels
             <integer> <character>        <integer>        <integer>
  [1]                1           *               37                0
  [2]               17           *               19                0
  [3]                6           *               13                0
  [4]                1           *               23                0
  [5]               30          -G                1                0
  [6]               39         -TC                1                0
  [7]               32           *                8                0
  [8]                3           *               18                0
  [9]               10           *               21                0
 [10]                1           *               28                0
 [11]                1           *                8                0
 [12]                5           *                3                0
 ---
 seqlengths:
  chr17
     NA
```

On the basis of current annotation, none of these indels are in exons:

```
> cdrt4txid <- as.character(elementMetadata(cdrt4txloc)$tx_id)
> cdrt4exloc <- exonsBy(txdb)[cdrt4txid]
> subsetByOverlaps(c17p.i, cdrt4exloc)

GRanges with 0 ranges and 11 elementMetadata values:
   seqnames    ranges strand | referenceBase consensusBase consensusQuality
      <Rle> <IRanges>  <Rle> |      <factor>      <factor>        <integer>
   snpQuality maxMappingQuality  coverage   alleleOne alleleOneSupport
    <integer>         <integer> <integer> <character>        <integer>
     alleleTwo alleleTwoSupport additionalIndels
   <character>        <integer>        <integer>
```

```
---
seqlengths:
 chr17
    NA
```

- Exercise: Write a function with parameters identifying a *GRanges* instance generated from a pileup, a gene symbol, a variant type, and a specification of feature scope, that reports on the variants present in the gene. Discuss infelicities of data structure in the code segment above that should be ameliorated to simplify solution of this exercise.

- Exercise: Whether or not you solve the previous exercise, characterize the variants in gene MYH3 for NA19240 in some concise way. It is advisable to focus on SNP; show that there are coding SNP present for this individual that are not identified in dbSNP.

- Exercise: Introduce and justify a mechanism for filtering variant reporting using quality information.

3.2 Compare Solexa calls with Sanger sequencing

- Setup: The 4 million phase II HapMap genotype calls for NA19240 are available to you in package hmyriB36. A selection confined to chromosome 6 is available in the *ind1KG* package.

```
> library(ind1KG)
> library(chopsticks)
> data(yri240_6)
> yri240_6$hm

A snp.matrix with  1 rows and  265955 columns
Row name:  NA19240
Col names:  rs4097465 ... rs4599694

> head(yri240_6$supp, 10)

           dbSNPalleles Assignment Chromosome Position Strand
rs4097465          G/T        G/T       chr6    37012      -
rs7754266          A/G        A/G       chr6    94609      +
rs9393087          C/T        C/T       chr6    94901      +
rs12192290         A/T        A/T       chr6    95272      +
rs11962658         A/C        A/C       chr6    96774      +
rs7742004          C/G        C/G       chr6    97749      +
rs2107722          G/T        G/T       chr6    98500      -
```

```
rs1929630          A/C        A/C       chr6    99536      +
rs12524398         C/G        C/G       chr6    99694      +
rs10484790         C/T        C/T       chr6    99750      -
```

- Exercise: Assess how many of the MAQ-based SNP calls using the chromosome 6 pileup data are found at dbSNP locations. Is the distribution of quality scores for variants identified at dbSNP locations similar to that of putatively de novo variants?

3.3 *de novo* SNPs in probes: effects on expression microarrays

Exercise: Acquire the probe sequences for the Illumina Human v1 expression array, perhaps by inverting the nuids found in the lumiHumanIDMapping metadata package.

```
> library(lumiHumanIDMapping)
> con <- lumiHumanIDMapping_dbconn()
> dbListTables(con)

 [1] "HUMANREF8_V3_0_R1_11282963_A_WGDASL" "HumanHT12_V3_0_R3_11283641_A"
 [3] "HumanHT12_V4_0_R1_15002873_B"        "HumanHT12_V4_0_R2_15002873_B"
 [5] "HumanHT12_V4_0_R2_15002873_B_WGDASL" "HumanRef8_V1"
 [7] "HumanRef8_V2_0_R2_11223162_A"        "HumanRef8_V2_0_R4_11223162_A"
 [9] "HumanRef8_V3_0_R0_11282963_A"        "HumanRef8_V3_0_R3_11282963_A"
[11] "HumanWG6_V1"                         "HumanWG6_V2_0_R2_11223189_A"
[13] "HumanWG6_V2_0_R4_11223189_A"         "HumanWG6_V2_11223189_B"
[15] "HumanWG6_V3_0_R3_11282955_A"         "metadata"
[17] "nuID_MappingInfo"

> dbGetQuery(con, "select * from HumanWG6_V1 limit 5")

    Search_key        Target    ProbeId   Accession Symbol               nuID
1        PLAC3 GI_23097300-A 0002360044 NM_021936.1  PLAC3 cn0dn1Sqdb0UHE4nEY
2         COG4 GI_21070955-A 0003940446 NM_015386.1   COG4 iklSlJ.eTo60t35XQE
3 GI_4505876-A  GI_4505876-A 0006420736 NM_000445.1  PLEC1 NBHBeFupql_azWVUMA
4        PTPRD GI_18860893-A 0002630279 NM_130393.1  PTPRD KcSlfQzU6Ld94lMSpE
5       HS6ST2 GI_27597081-A 0003120162 NM_147174.2 HS6ST2 ZeMrPvoCSjgl4lLoAk
```

Determine the genomic positions of all probes interrogating genes on chromosome 17 using *Biostrings* `matchPDict` against the consensus genomic sequence for chromosome 17. Find all probes (on chr17) corresponding to sequence for which NA19240 is found by MAQ to harbor a variant (use the pileup noted previously). We will call these probes "associated with sequence variants". Compute expression Z-scores for expression levels obtained for NA19240 using mean and standard deviation based on log expression for the

89 individuals in hmyriB36 excluding NA19240. Can the distribution of expression Z-scores for probes associated with sequence variants be distinguished from the distribution of expression Z-scores for probes not associated with sequence variants.

Extra credit extension: Some probes define sequence associated with splice junctions. These 50mers will not align to consensus genomic sequence, but will align once introns are removed. Can you identify probes associated with splice junctions that are also associated with sequence variants? Does the expression Z-score for splice-junction-associated probes differ in distribution from the general distribution of expression Z-scores?

Additional exercises. Retrieve the SOLiD or 454-based short read archives for NA19240 and check the consistency of conclusions obtained in prior Solexa-based exercises with results based on these platforms.

4 Session information

```
> sessionInfo()

R version 2.14.2 (2012-02-29)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] splines   stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] lumiHumanIDMapping_1.10.0
 [2] lumi_2.6.0
 [3] nleqslv_1.9.3
 [4] methylumi_2.0.13
 [5] SNPlocs.Hsapiens.dbSNP.20090506_0.99.6
 [6] org.Hs.eg.db_2.6.4
 [7] RSQLite_0.11.1
 [8] DBI_0.2-5
 [9] chopsticks_1.18.6
[10] survival_2.36-12
```

```
[11] ind1KG_0.1.12
[12] TxDb.Hsapiens.UCSC.hg18.knownGene_2.6.2
[13] GenomicFeatures_1.6.8
[14] AnnotationDbi_1.16.19
[15] Biobase_2.14.0
[16] Rsamtools_1.6.3
[17] Biostrings_2.22.0
[18] GenomicRanges_1.6.7
[19] IRanges_1.12.6

loaded via a namespace (and not attached):
 [1] BSgenome_1.22.0       BiocInstaller_1.2.1   KernSmooth_2.23-7
 [4] MASS_7.3-17           Matrix_1.0-4          RCurl_1.91-1
 [7] XML_3.9-4             affy_1.32.1           affyio_1.22.0
[10] annotate_1.32.1       biomaRt_2.10.0        bitops_1.0-4.1
[13] grid_2.14.2           hdrcde_2.15           lattice_0.20-0
[16] mgcv_1.7-13           nlme_3.1-103          preprocessCore_1.16.0
[19] rtracklayer_1.14.4    tools_2.14.2          xtable_1.7-0
[22] zlibbioc_1.0.1
```