CNV calling, and a set of other useful notes

Vincent Plagnol

Inivata- Head of Computational Biology
UCL- Reader in Statistical Genetics

Longer reads exist if you need them

```
Query= ch442_file7_twodirections
         (1880 letters)

>Dscam_9.11_6.27_4.7
          Length = 2207

 Score = 1552
 Identities = 1750/1878 (93%), Gaps = 101/1878 (5%)
 Strand = Minus / Plus

Query: 1871 TCGGATCCATTATCTCCCGGGACGTCCATGTGCGAGCCG-TTGTCAATCAGTTCTATGAG 1813
Sbjct: 307  TCGGATCCATTATCTCCCGGGACGTCCATGTGCGAGCCGGTTGTCAATCAGTTCTATGAG 366

Query: 1812 GCGGAGATTATGACCG-GTATGTTATCATAGGGTAATGCTGCCGTTCTGAAGTGTTCCAT 1754
Sbjct: 367  GCGGAGATTATGACCGAGTATGTTATCAG-GGGTAATGCTGCCGTTCTGAAGTGTTCCAT 425

Query: 1753 CCCGTCGTTCGTGAGCTCCATTTGTTCGAGTTGAGTCCTGGATCGATGATGAGGGCAATG 1694
Sbjct: 426  CCCGTCGTTCGTG-GCTGATTTTGTTCGAGTTGAGTCCTGGATCGATGATGAGGGCAATG 484

Query: 1693 TCTTGTCCTTTTCGGACAATTACG-ATGGAAAGTACCTGGTATTGCCCTCTGGAGAGCTG 1635
Sbjct: 485  TCTTGTCCTTTTCGGACAATTACGGATGGAAAGTACCTGGTATTGCCCTCTGGAGAGCTG 544

Query: 1634 CACATCCGTGAGGTTG-ACCCGAGGACG-ATACAAGAGCTACCAGTGCCGAACCAAACAT 1577
Sbjct: 545  CACATCCGTGAGGTTGGACCCGAGGACGGATACAAGAGCTACCAGTGCCGAACCAAACAT 604

Query: 1576 CGTCTGACCGGAGAAACCCGATCCTAAGTGCCACAAAAGGACGATTGGTCATCACAG-AA 1518
Sbjct: 605  CGTCTGACCGGAGAAACCCGAT--TAAGTGCCACAAAAGGACGATTGGTCATCACAGGAA 662

Query: 1517 CCTGTGG-CAGTGTTGCGCCTAAGATCAGCGTAGAAAATCGTCTGAAAA-TGCAGAGGCT 1460
Sbjct: 663  CCTGTGGGCAGTGTTGCGCCTAAGATCAGCGTAGAAAATCGTCTGAAAAATGCAGAGGCT 722

Query: 1459 CGAATTTGGTTCCA--TTCACCCTCTTT-GCCCTGGTCAAGCCTATCCAGTACCCTTCTT 1403
Sbjct: 723  CGAATT-GGTTCCACCTTCACCCTCTTTTGCCCTGGTCAAGCCTATCCAGTACCCTTCTT 781

Query: 1402 TAG-ATGGTACA-GTTCATCGAAGGAACAACCCGTAAA--AGCTCTTGGTGCTCTAAATG 1347
Sbjct: 782  TAGGATGGTACAAGTTCATCGAAGGAACAACCCGTAAACAAGCTGT-GGTGCT--AAATG 838

Query: 1346 AC---GTGAAGCAGGTCTCTGGTACTCTCGGGCCCTTATTAAGGACGCAGTAGGTCGAGG 1290
Sbjct: 839  ACCGTGTGAAGCAGGTCTCTGGTACTCTCA-----TTATTAAGGACGCAGTAG-TCGAGG 892

Query: 1289 ACTCCGGAAAGGTACCTGTGCGTGGTAAACAATTCCGTGGGAGGCGAAAGCGTTGAGACT 1230
Sbjct: 893  ACTCCGGAAAA-TACCTGTGCGTGGTGAACAATTCCGTGGGAGGCGAAAGCGTTGAGACT 951

Query: 1229 GTGCTTACCGTCACCTGCCCCC-TATCCGCCTAAGATCGATCCTCCCACACAGACAGTTG 1171
Sbjct: 952  GTGCTTACCGTCAC-TGCCCCCCTATCCGC-TAAGATCGATCCTCCCACACAGACAGTTG 1009

Query: 1170 ACTTCGGACGCCCCGCCGTATTCACT-GCCAGTACACTGGAAATCCCATCAAGACCGTTA 1112
Sbjct: 1010 ACTTCGGACGCCCCGCCGTATTCACTTGCCAGTACACTGGAAATCCCATCAAGACCGTTA 1069

Query: 1111 GCTGGGATGAAGGACGGCAAGGCTATCGGACA--GCGAACCCGTGCTGCGCATCGAATCC 1054
Sbjct: 1070 GCTGG-ATGAAGGACGGCAAGGCTATCGGACACAGCGAACCCGTGCTGCGCATCGAATCC 1128

Query: 1053 GTGAAGAAGGAGGACAAGG-CATGTACCAGTGCTTCGTAAGAAACGACCAGGAATTACGG 995
Sbjct: 1129 GTGAAGAAGGAGGACAAGGGCATGTACCAGTGCTTCGTAAGAAACGACCAGGAAT--CGG 1186

Query: 994  CGGAGGCGTAGTGCTGAGCTGAA--TGGGAGGCCGTTTCGATCCCCC-GTCATCCGACAG 938
Sbjct: 1187 CGGAGGCG-AGTGCTGAGCTGAAGCTCGGAGGCCGTTTCGATCCCCCCGTCATCCGACAG 1245

Query: 937  GCCTTCCAGGAAGAGACCATGGAACCCGGACCAAGTGTATTCCTCAAGTGCGTCCCCGGC 878
Sbjct: 1246 GCCTTCCAGGAAGAGACCATGGAACCCGGACCAAGTGTATTCCTCAAGTGCGTCGCCGGC 1305

Query: 877  GGAAACCCCACGCCCGAAATT-CCTGG-AACTCGACGGCACAAGAAAATCGCCGCCAACA 820
Sbjct: 1306 GGAAACCCCACGCCCGAAATTTCCTGGGAACTCGACGGCA--AGAAAATCGCC---AACA 1360

Query: 819  ACGATCGCTATCAGGTCGGACAGTATGTATGACGGTCAACGGAGATGTGGTTTCCTA-CT 761
Sbjct: 1361 ACGATCGCTATCAGGTCGGACAGTATGT--GACGGTCAACGGAGATGTGGTTTCCTATCT 1418

Query: 760  CAAACATAACCTCGGTCCACGCCAACGATGGC--T-TCTA--AGTGCA--GCTCAAGAGC 708
Sbjct: 1419 GAA-CATAACCTCGGTCCACGCCAACGATGGCGGTCTCTACAAGTGCATAGC-CAAGAGC 1476

Query: 707  AAAGTGG-CGTGGCCGAACACTGGGGTACATTGAATGTATATGGACTGCC-TACATCCGG 650
Sbjct: 1477 AAAGTGGGCGTGGCCGAACACTCGGCCAAGTTGAATGTATATGGACTGCCCTACATCCGG 1536

Query: 649  CAATGGGAGAAGAAGGCAATTGTTGCTGGAGAAACCCTGATTGTCACCACATGTCCTGTA 590
Sbjct: 1537 CAAATGGAGAAGAAGGCAATTGTTGCTGGAGAAACCCTGATTGTCAC---ATGTCCTGTA 1593

Query: 589  GCTGGATACCCCATCGAGCATTCGATTGTCTGGGAGCGAG-ATAACCGTGCCCTGCCCAT 531
Sbjct: 1594 GCTGGATACCCCATCGA---TTCGATTGTCTGGGAGCGAGGATAACCGTGCCCTGCCCAT 1650

Query: 530  CAACCGCAAACAGAAGGTCTTCCCCAACGGAACGCTGATTATCGAGAATGTGGAACGAGA 471
Sbjct: 1651 CAACCGCAAACAGAAGGTCTTCCCCAACGGAACGCTGATTATCGAGAATGTGGAACG-GA 1709

Query: 470  ACTCAGACCAAGCGACCTACA--TGCGTTGCCAAGAATCAGGAAGGAT-CTCTGC-CGCG 415
Sbjct: 1710 ACTCAGACCAAGCGACCTACACTTGCGTTGCCAAGAATCAGGAAGGATACTCTGCTCGAG 1769

Query: 414  GATCTCTGGAAGTGCAAGTCATGG-TGCCTCCACAAATTCGACCCTTTGACTTCGGGGAC 356
Sbjct: 1770 GATCTCTGGAAGTGCAAGTCATGGGTGCCTCCACAAATTCGACCCTTTGACTTCGGCGAC 1829

Query: 355  GAGGCTTCCAACTCAGGCGAAACCATGGGCATTCCCGTGCACGGTGATCAA-GGGGATCT 297
Sbjct: 1830 GAGGCTTCCAACTCAGGCGAAACCATGGGCATTC-AGTGCACGGTGATCAAGGGGGATCT 1888

Query: 296  GCCCATCAATATTACGT-GCTTCACCAACAATCACACCCTGAACAGTGGTGACCTAGAC- 239
Sbjct: 1889 GCCCATCAATATTACGTGGGTTC-TCAACAATCACACCCTGAACAGTGGTGACCTAGACG 1947

Query: 238  --GTGATCGGTCGAATGTCCAGCAAGTCGAGCACCCCTGAATATCGACTACATAGCCGCC 181
Sbjct: 1948 TGGTGATCGGTCGAATGTCCAGCAAGTCGAGCA-CCCTGAATATCGACTACATAGCCGCC 2006

Query: 180  GAGCACCGTGCGGTCTACACCTGTCT-GCTCGCAATC-GGC-GGCGAGAGCAGCTACAGT 124
Sbjct: 2007 GAGCACCGTGGGGTCTACACCTGTCTGGCTCGCAATCAGGCGGGCGAGAGCAGCTACAGT 2066

Query: 123  GCTGAGCAGTTGAAAGTCAAC-GTACCTCCCAGATGGATCCTC--GCTCACCGATGACTA 67
Sbjct: 2067 GCTG---AGTTGAAAGTCAACGGTACCTCCCAGATGGATCCTCGAGCCCACCGATAA--G 2121

Query: 66   GCCTTCG-CCAGGGATCC 50
Sbjct: 2122 GCCTTCGCCCAGGGATCC 2139
```

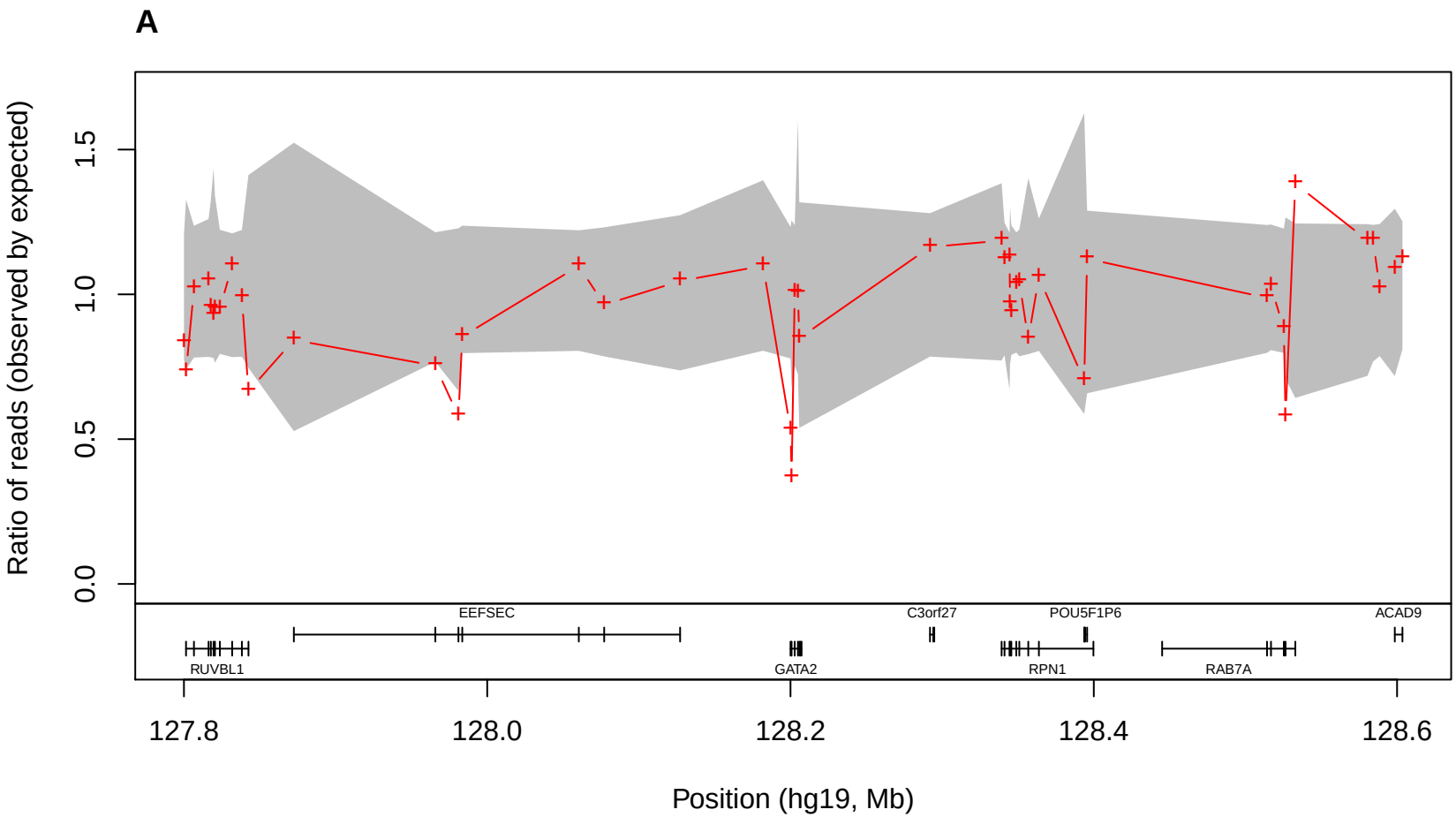
BEDtools: your swiss army knife for all issues

- BEDtools is one of the most widely used tools in bioinformatics.
- It does tons of things, and while many look trivial, together they are very impressive.
- Well worth looking at what it can do, because it may solve many practical questions.
    - In R, most routines are implemented within `GenomicRanges`, a bioconductor package.

Calling CNVs

- There are many tools to call CNVs from sequence data, and I think you have already covered some of them.
- One option is read depth: excess of reads mark a duplication, too few reads mark deletions.
- But there are also specific read patterns, like reads mapping further apart than they should, that can mark a deletion.
- Split read is another way to go about it.

An example of read depth based call in GATA2



Copy number variant analysis