Fig S5: SNPdip Figure For "Supplement"

Chr1-unfiltered

March 22, 2018

Contents

1 Intro

This document just generates the "SNPdip" figure for the Supplement; it was once ED 3, later ED 4; it is currently S5. Much more explanation and analysis of this can be found in SNPdip.rnw (from which most of this was lifted).

2 Preliminaries

Load utility R code; do setup:

```
source('../../../R/wlr.R') # load util code; path relative this folder or sibling in scripts/larrys

## Running as: ruzzo @ macreg15026.dyn.cs.washington.edu; SVN Id, I miss you.  $Id: wlr.R  2017-07-21 or later $

setup.my.wd('paperfigs')    # set working dir; UPDATE if this file moves, or if COPY/PASTE to new file
setup.my.knitr()
figdir <- 'FigS5-SNPdip-figs/'
generic.setup(figdir)             # Create figs dir etc., if needed.
```

Load the tables. By default, this will just build/cache/load the Chr1 subset. Full genome analysis should be possible, but I doubt the extra data will reveal anything new.

```
snp.tables.chr1 <- load.snp.tables(use.chr1.tables=TRUE, data.name='full.tables.01.26.14') # see wlr.R for paths

# Loading ../00common/mycache/snp.tables.chr1.unqfiltered.rda ...Loaded.
```

A LaTeX hack: I want which.snp.tables info in doc title/page headers, but it is unknown until now, so the following writes a command definition `\whichsnptables` into the .aux file, which is read during the *next* LaTeX run, when `\begin{document}` is processed:

```
\makeatletter
\immediate\write\@auxout{\noexpand\gdef\noexpand\whichsnptables{Chr1-unfiltered}}
\makeatother
```

3 Average Coverage Drops Around SNPs

Basic story is that average coverage drops around SNPs. We presume the reason is that the short read mapping software tolerates only a small number of mismatches to the reference genome. A correct read across an alternate allele has, by definition, at least one mismatch, so additional read errors and/or low quality positions are more likely to push the read below the mapper's alignment threshold, resulting in lower coverage. `dip.summary` below calculates the desired summary statistics, and `hilodip` plots them. [NOTE: this function is slow, taking 10-15 minutes per call to process Chr1. To facilitate debugging of layout, plot formats, derivative analyses, etc., all calls should include d.r.e and d.r.name parameters; then flipping T/F below will toggle *all* between a short 10k test case and the full Chr 1 analysis.]

```
if(T){                  # small example for testing
  d.r.name <- 'tenktest'    # dip.region.name
  d.r.e <- 10000            # dip.region.end
} else {                # all of Chr 1
  d.r.name <- 'Chr1'
  d.r.e <- length(snp.tables.chr1[[1]]$Cov)
}
```

```
cachet('dip.sum', dip.summary(d.r.name, dip.region.end=d.r.e, snp.tables=snp.tables.chr1))

# Loading... dip.sum
```

```
for(st in 1:7){
  pdf(paste(figdir, 'snpdip-chr1-', names(snp.tables.chr1)[st], '.pdf',sep=''), height=5, width=10);
  showdip(st, dip.sum,c(T,T,F,F,F,T,'both'))
  dev.off()
}
```

The story is quite solid—although there is considerable variability when the number of SNPs is small (e.g., Gyre), when averaged over sufficiently many SNP positions, the average coverage is reasonably flat for positions more than 25 bases away from a SNP (vertical grey lines), but drops linearly as one approaches the SNP. These boundaries and the linear drop are also as expected: Most sequencing was done with 25bp reads (actually 25 x 2 mate paired reads), so positions more than 25 bases from a SNP are unaffected by it, but a read covering a position $d < 25$ positions away has a chance of $\approx (25 - d)/25$ of *also* covering the SNP, and as explained above, there is a bias against aligning those reads. As further confirmation of this model, I believe that we included additional sequence data for both Italy and NY—35bp (unpaired) reads—and plots for both show a slightly wider "vee" with a slightly shallower slope for the last $\approx$10 positions, reflecting the portion of coverage derived from the longer reads.

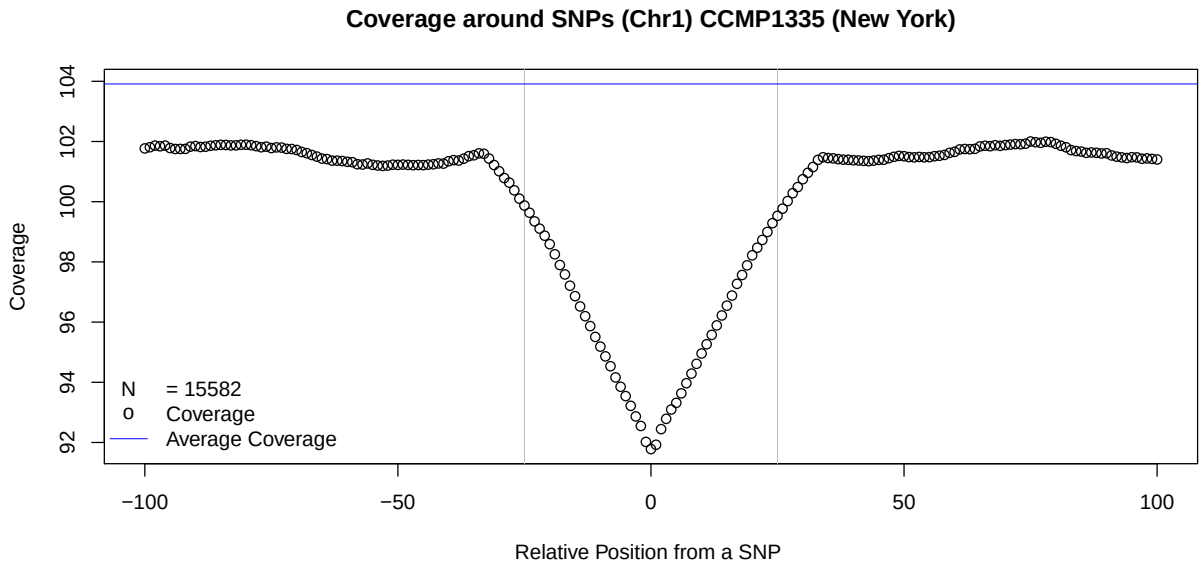
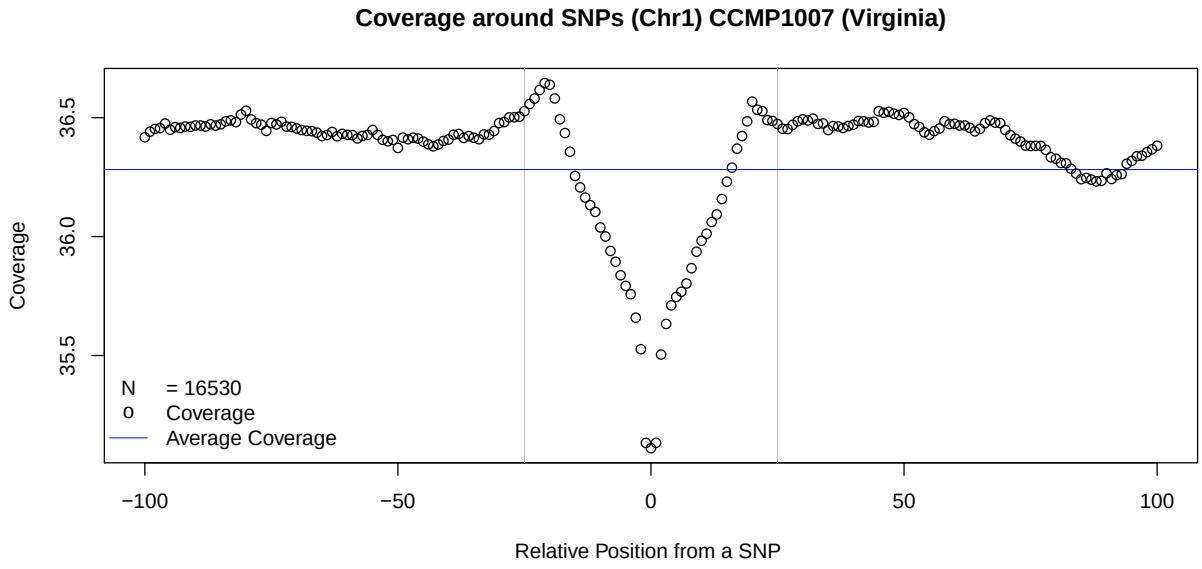
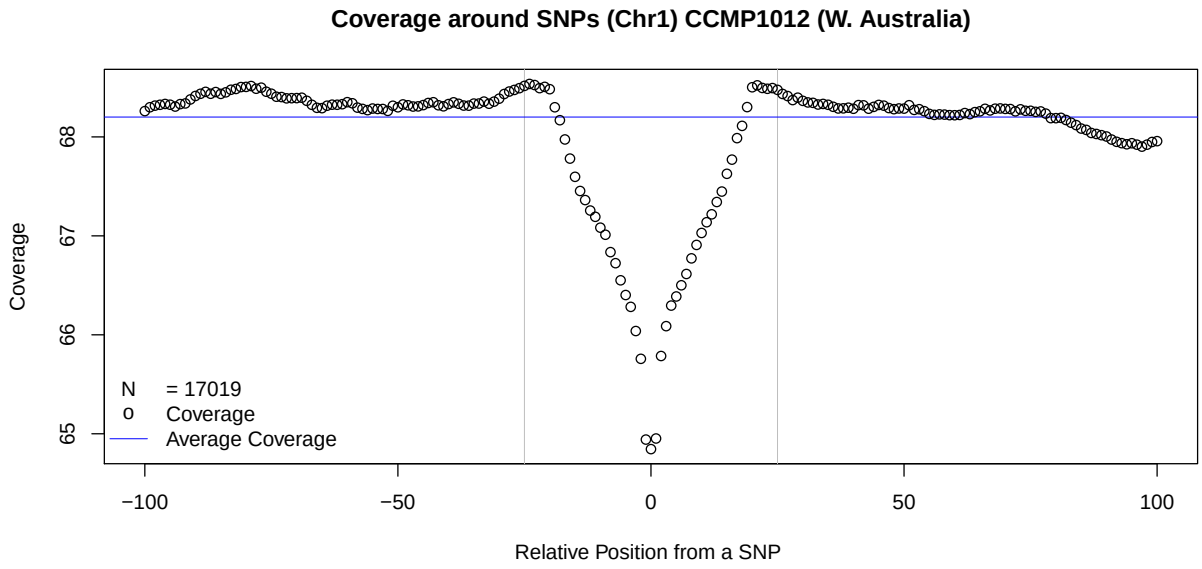
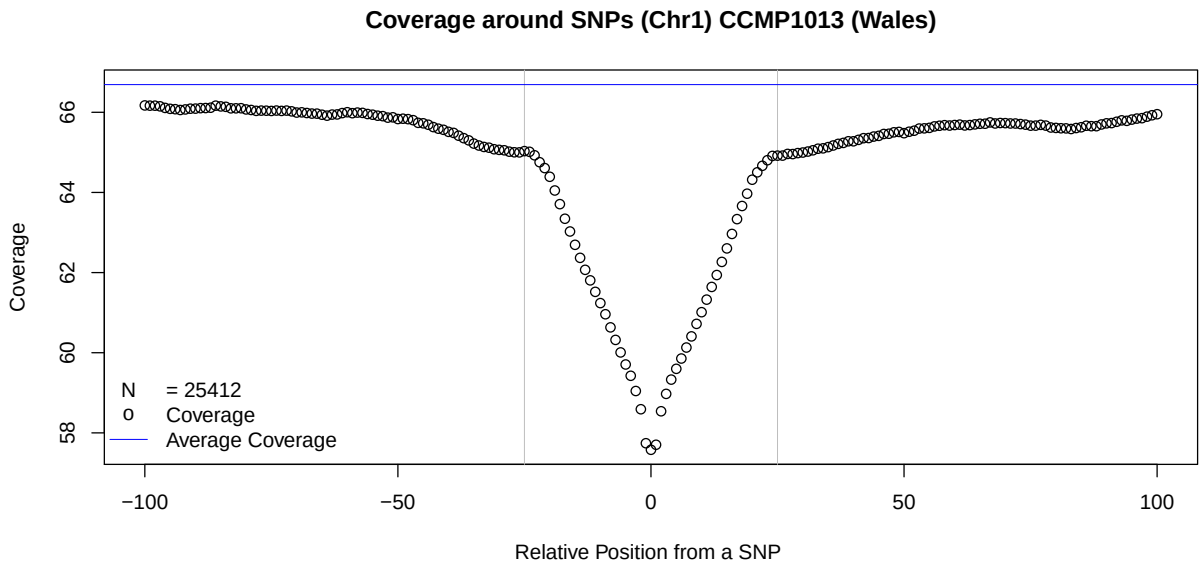


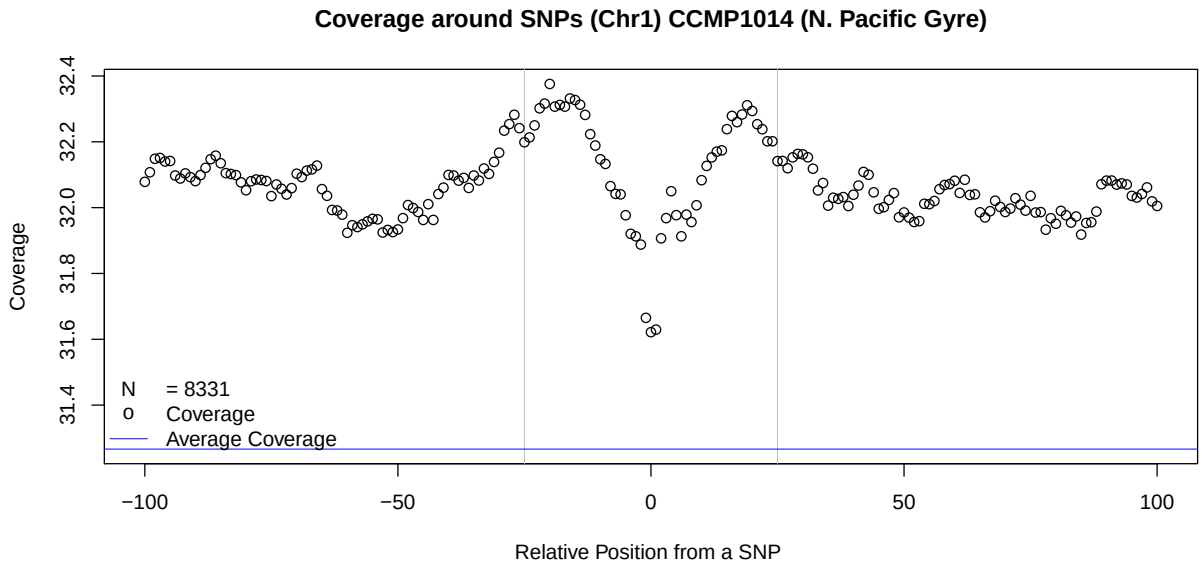
Figure 1: **Suggested Caption: Read depth near SNPS.** Plotted is the read depth within ± 100 nucleotides of single nucleotide polymorphisms (SNPs) identified by SAMtools, averaged over all 15582 SNPs called on Chromosome 1 of CCMP 1335. X-axis: distance from SNP (bp). Y-axis: read depth. Horizontal blue line: average read depth across Chromosome 1. Vertical grey lines: ± 25bp from SNP. [[FOLLOWING TEXT IS PROBABLY BETTER IN THE ACCOMPANYING PROSE THAN IN THE CAPTION, BUT SOMEWHERE, SAY SOMETHING LIKE: CCMP1335 was sequenced with a mixture of 25bp mate-paired reads and 36bp single-end reads (TONY: ARE THESE NUMBERS CORRECT?). The vertical grey lines in the figure mark the range in which the 25bp reads could simultaneously include the indicated position and the central SNP. The probability that a randomly placed read covering a position d base pairs away from the central SNP will *not* cover the SNP rises linearly with $|d|$, which presumably explains the observed linear increase in coverage over the $\pm$read length interval.]]

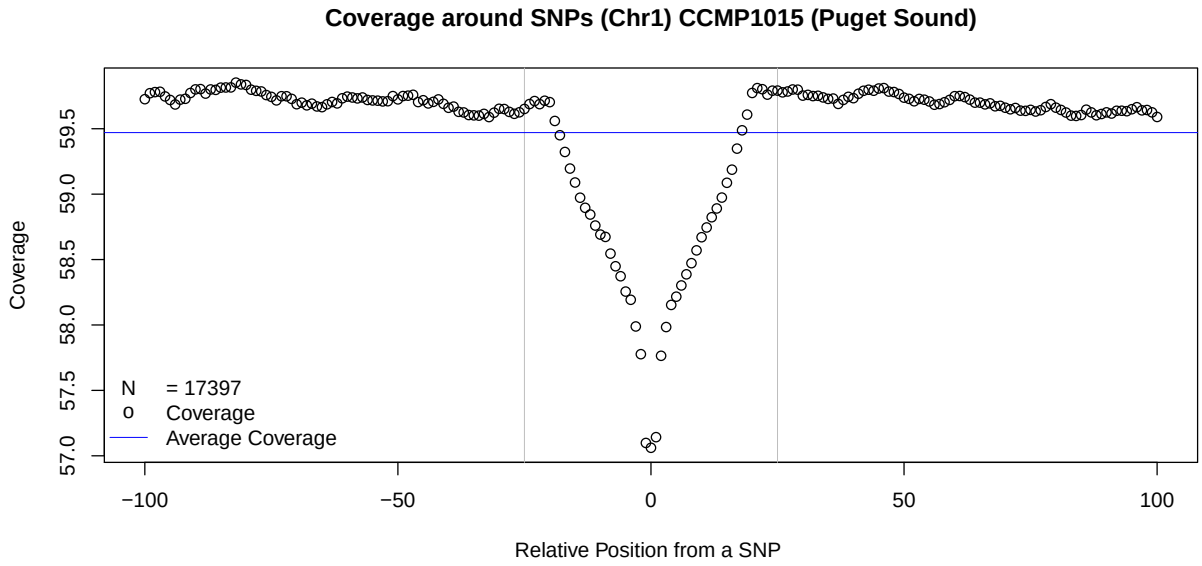
Here are the others for comparison; pick whichever.

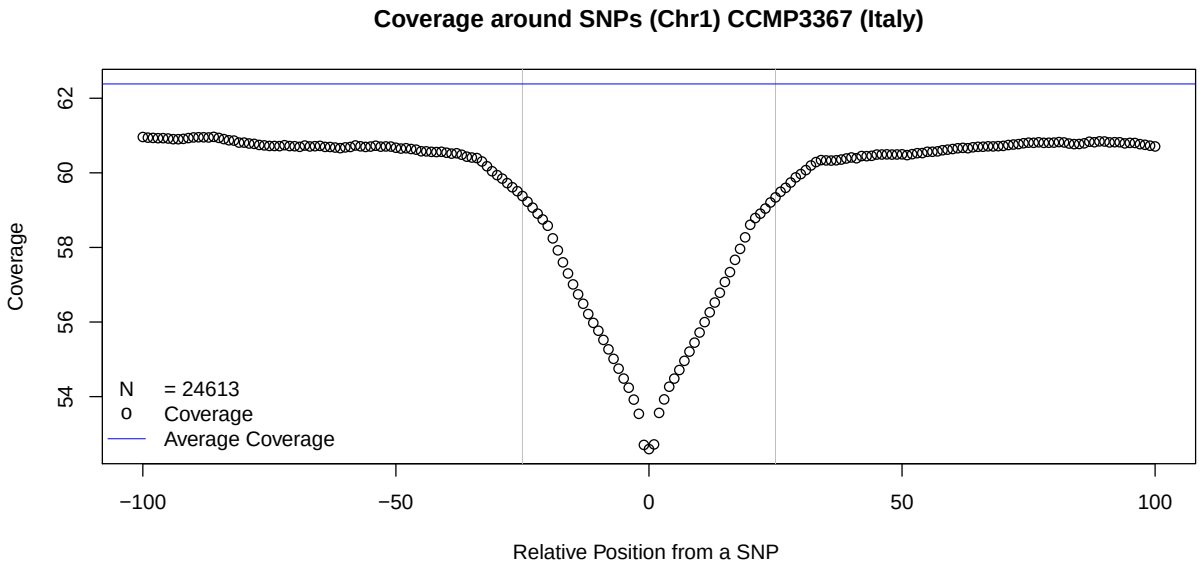



Coverage around SNPs (Chr1) CCMP1012 (W. Australia)
Coverage
65
66
67
68
N = 17019
o Coverage
Average Coverage
−100
−50
0
50
100
Relative Position from a SNP


Coverage around SNPs (Chr1) CCMP1013 (Wales)
Coverage
58
60
62
64
66
N = 25412
o Coverage
Average Coverage
−100
−50
0
50
100
Relative Position from a SNP


Coverage around SNPs (Chr1) CCMP1014 (N. Pacific Gyre)
Coverage
32.4
32.2
32.0
31.8
31.6
31.4
N = 8331
o Coverage
Average Coverage
−100
−50
0
50
100
Relative Position from a SNP


Coverage around SNPs (Chr1) CCMP1015 (Puget Sound)
Coverage
59.5
59.0
58.5
58.0
57.5
57.0
N = 17397
o Coverage
Average Coverage
−100
−50
0
50
100
Relative Position from a SNP



SVN ID I miss you Id: FigS5-SNPdip.rnw 2017-05-29 or later ruzzo