Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Hacein-Bey-Abina S, Pai S-Y, Gaspar HB, et al. A modified γ-retrovirus vector for X-linked severe combined immunodeficiency. N Engl J Med 2014;371:1407-17. DOI: 10.1056/NEJMoa1404588

Supplementary Appendix

Table of Contents:

Supplemental methods

Data collection and manuscript preparation

The study was designed by Adrian Thrasher. The data were gathered centrally using INFORM database. The clinical data were analyzed by Dr. Wendy London from a biostatistical standpoint and Frederic Bushman for integration site analysis with Salima Hacein-Bey-Abina, Sung-Yun Pai, Luigi Notarangelo, Alexandra Filipovich, Donald Kohn, Bobby Gaspar, Marina Cavazzana, and David Williams also participating in analysis of data and vouching for the data. Sung-Yun Pai, Salima Hacein-Bey Abina, and David Williams wrote the paper with major edits by Alain Fischer, Luigi Notarangelo, Marina Cavazzana, Alexandra Filipovich, Don Kohn, Frederic Bushman, and Adrian Thrasher. All co-authors reviewed the manuscript. All co-authors agreed to publish the paper.

David Williams served as IND holder and trial sponsor in the United States. The agreement to share data between the sponsor of the trial and both US and European investigators was included in the Informed Consent Form stating:

> Maintaining confidentiality of your child's information is very important to us. All research data pertaining to your child will only be available to people directly involved in the research study and will be kept in electronic form in password protected files in an area of the hospital computer accessible only to research staff. Paper records will be kept in locked files, cabinets or offices. Information will be released to research staff at Children's Hospital and Dana-Farber Cancer Institute to the sponsor of the study (Dr. David Williams) and his staff, to the funding agency (NIAID/DAIT) and to research staff at the other participating institutions in this trial (Great Ormond Street Hospital, Hospital Necker Enfants Malades, Cincinnati Children's Hospital Medical Center, Mattel Children's

Hospital UCLA). In addition, relevant information will be released to federal and state agencies that have authority over the research, namely, the Food and Drug Administration and the National Institutes of Health Recombinant DNA Advisory Committee.

Determination of Vector Copy Number (VCN) by real-time PCR (qPCR)

Genomic DNA was extracted from $CD34^+$ liquid cultures, whole peripheral blood, Ficoll-purified blood mononuclear cells and peripheral blood subpopulations with the DNeasy Blood and Tissue kit (Qiagen, Germantown, MD). Transgene sequences were detected by a duplex qPCR (FAM/VIC TaqMan® assay) for simultaneous detection of γc transgene and *APOB*. Standard curves were obtained by serial dilutions of a plasmid (pSRS11-gC/pre/ApoB) containing γc transgene and *APOB* sequences. Genomic DNA from a clone containing 1 copy γc/cell was included as a control. The number of integrated VCN per cell (i.e diploid genome) was determined by multiplying the ratio γc/APOB by two.

Primers and Probes sequences used: ApoB (Fwd TGAAGGTGGAGGACATTCCTCTA; Rev CTGGAATTGCGATTTCTGGTAA; VIC-CGAGAATCACCCTGCCAGACTTCCGT-TAMRA); γc (Fwd TGCTAAAACTGCAGAATCTGGT; Rev AGCTGGGATTCACTCAGTTTG; 6FAM- CCTGGGCTCCAGAGAACCTAACA-TAMRA)

Integration site sequence acquisition and analysis

Overview: The DNA sequencing methods used in the integration site analysis differed among trials--the London SCID MFG-gamma-c trial samples were analyzed using cleavage of genomic DNA with restriction enzymes and the Sanger method; the Paris SCID MFG-gamma-c trial samples were analyzed using cleavage of genomic DNA

with restriction enzymes and 454/Roche pyrosequencing; and the samples for the SIN-gamma-c were analyzed by random cleavage with fragmentase and 454/Roche pyrosequencing as detailed below.

DNA was purified from samples of sorted blood cells and analyzed by ligation-mediated PCR. Samples were as follows:

Subject	Number of DNA samples analyzed for integration site distributions
Pt #1	16
Pt #2	28
Pt #3	19
Pt #5	24
Pt #6	11
Pt #7	5
Pt #8	14
Pt #9	8

DNA samples were cleaved at random locations, DNA adaptors ligated to the cleaved ends, junction fragments amplified by PCR using primers annealing to the adaptor and the vector DNA end, and libraries analyzed by pyrosequencing [1]. As a control, human DNA samples lacking integrated vectors were analyzed in parallel and shown to be free of spurious integration sites (n=30 controls), documenting absence of PCR contamination.

Sample work up and pyrosequencing: Prior to library preparation, genomic DNA (gDNA) was repurified using Agencourt AMPure XP Beads at a 1:1 bead to DNA ratio (v/v). The gDNA was then quantified with the Quant-iT Picogreen dsDNA Assay Kit and fragmented with NEBNext dsDNA Fragmentase. Fragmentation reactions were prepared according to the manufacturer's protocol and incubated for 50 minutes at 37°C,

yielding fragments of 100-500bp. Fragmented DNA was purified using the QIAquick PCR Purification Kit. End-repair and 3' adenylation were performed with the NEB Quick Blunting Kit and NEB Klenow Fragment (3'→5' exo-) respectively. Adenylation reactions were purified using the QIAquick PCR Purification Kit.

Seventy unique DNA adaptors were annealed by mixing equimolar amounts of two oligonucleotides and slowly cooling the mixture from 95°C to 4°C over 2.5 hours. Each adaptor contained a unique ssDNA primer landing site template and a common dsDNA sequence containing a 5' T-overhang. The annealed adaptors were ligated, one per sample, to DNA fragments containing 3' A-overhangs using T4 DNA ligase. Two rounds of touchdown PCR were used to recover the DNA junctions of the vector and human host. Each round of PCR utilized one primer designed to anneal to the viral LTR and one primer designed to anneal to the unique adaptor sequence. The second round PCR primer on the LTR side contained a DNA bar code that designated the patient, time point, and cell type analyzed. Adaptor-to-adaptor amplification was suppressed in the initial PCR reaction by using a 3' Amino block on the common strand of the adaptor. PCR crossover was suppressed in all PCR reactions by handling each reaction individually in a PCR hood and using unique adaptors and adaptor priming sites in each PCR reaction [2].

The initial PCR reactions were prepared in quadruplicate along with one negative control per set of samples. 25uL PCR reactions were prepared with the Advantage 2 PCR Kit using the concentrations of reagents recommend by the manufacturer. The initial PCR reactions were cycled with the following parameters: 1x 1' 94˚C; 5x 2'' 94˚C, 1' 72˚C; 25x 2'' 94˚C, 1' 70˚C; 1x 4' 72˚C; 4˚C hold. PCR products were diluted 1:100 in preparation for the second (nested) PCR. The nested PCR adaptor primer included the Roche 454 A primer and the nested PCR LTR primer included an 8nt barcode (one per replicate) and the Roche 454 B primer. The nested PCR reactions were cycled with the

same conditions as the initial PCR except that the second touchdown step was repeated 20 times instead of 25. Nested PCR products from the four replicates for each sample were pooled and size-separated on a 0.8% agarose gel stained with SYBR Safe DNA Gel Stain. DNA was extracted from gel slices containing fragments of 300-500bp using the QIAquick Gel Extraction Kit. Samples were further purified using Agencourt AMPure XP Beads at a 0.9:1 bead to sample ratio (v/v) to ensure removal of small amplicons and primers. Sample concentrations were measured using the Quant-iT Picogreen dsDNA Assay Kit. Samples were pooled equally by mass and sequenced from the B direction on a 454 GS Junior System. 500ng of naïve HEK-293T DNA was processed in parallel to each set of samples to serve as a biological negative control, and consistently lacked spurious integration sites.

Bioinformatic analysis. All sequence reads were required to show the correct pairing of unique adaptor and bar code. Sequences were quality controlled by requiring a 98% match to the human genome, and the match was required to begin within three bases of the vector LTR edge. Sequences were aligned to the hg18 draft of the human genome, and beginning and ending coordinates stored for each alignment. The ROC area method was used to assess the relationship of integration site distributions to genomic annotation or epigenetic marks by comparison to random distributions [3]. A detailed description of the statistical methods used can be found in [4]. Analysis of gene ontology showed similar association of sites from all trials with major gene categories such as “phosphoprotein” and “alternative splicing”. Abundance of cell clones was assessed by counting the number of different random break points that gave rise to capturing each integration site, thereby taking advantage of covalent DNA marks introduced into the DNA prior to PCR amplification. Abundance was analyzed statistically using the SonicLength method, and population sizes estimated using the Chao estimator with jackknife correction as described [5]. Analysis of integration site

clusters was carried out using scan statistics [6], which allows statistical analysis without making assumptions about the genomic length of clusters or the number of integration sites involved (see the attached Supplementary Report for details). Curated cancer gene lists used in this study are described here [7] and available here (http://www.bushmanlab.org/links/genelists).

The human lymphoid cancer gene list contained the following genes:

geneID	symbol	geneName
25	ABL1	c-abl oncogene 1, non-receptor tyrosine kinase
596	BCL2	B-cell CLL/lymphoma 2
53335	BCL11A	B-cell CLL/lymphoma 11A (zinc finger protein)
64919	BCL11B	B-cell CLL/lymphoma 11B (zinc finger protein)
613	BCR	breakpoint cluster region
648	BMI1	BMI1 polycomb ring finger oncogene
6046	BRD2	bromodomain containing 2
595	CCND1	cyclin D1
894	CCND2	cyclin D2
1045	CDX2	caudal type homeobox 2
2120	ETV6	ets variant 6
3717	JAK2	Janus kinase 2
3727	JUND	jun D proto-oncogene
1316	KLF6	Kruppel-like factor 6
3932	LCK	lymphocyte-specific protein tyrosine kinase
4004	LMO1	LIM domain only 1 (rhombotin 1)
4005	LMO2	LIM domain only 2 (rhombotin-like 1)
4066	LYL1	lymphoblastic leukemia derived sequence 1
4297	MLL	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)
8028	MLLT10	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 10
4515	MTCP1	mature T-cell proliferation 1
4609	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
4791	NFKB2	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)
4851	NOTCH1	notch 1
4928	NUP98	nucleoporin 98kDa
10215	OLIG2	oligodendrocyte lineage transcription factor 2
5087	PBX1	pre-B-cell leukemia homeobox 1
8301	PICALM	phosphatidylinositol binding clathrin assembly protein
5910	RAP1GDS1	RAP1, GTP-GDP dissociation stimulator 1
861	RUNX1	runt-related transcription factor 1
6491	STIL	SCL/TAL1 interrupting locus

6886	TAL1	T cell acute lymphocytic leukemia 1
6887	TAL2	T cell acute lymphocytic leukemia 2
154215	NKAIN2	Na+/K+ transporting ATPase interacting 2
6929	TCF3	transcription factor 3
8115	TCL1A	T-cell leukemia/lymphoma 1A
3195	TLX1	T-cell leukemia homeobox 1
30012	TLX3	T-cell leukemia homeobox 13

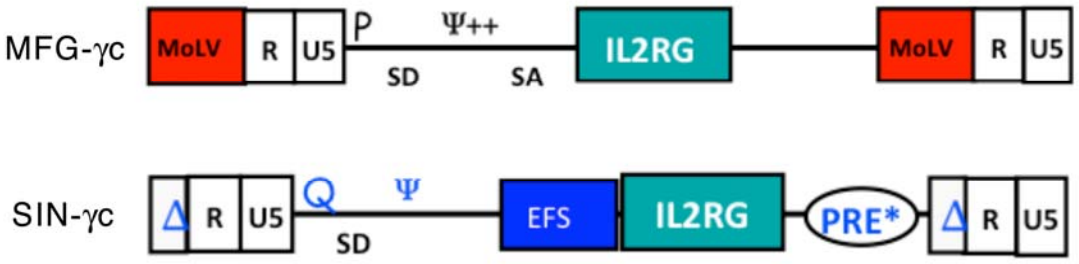
A detailed study of integration site sharing between cell types will be reported elsewhere. All integration site data sets will be deposited at the NCBI Sequence Read Archive upon acceptance of this paper for publication.

Supplemental Figures

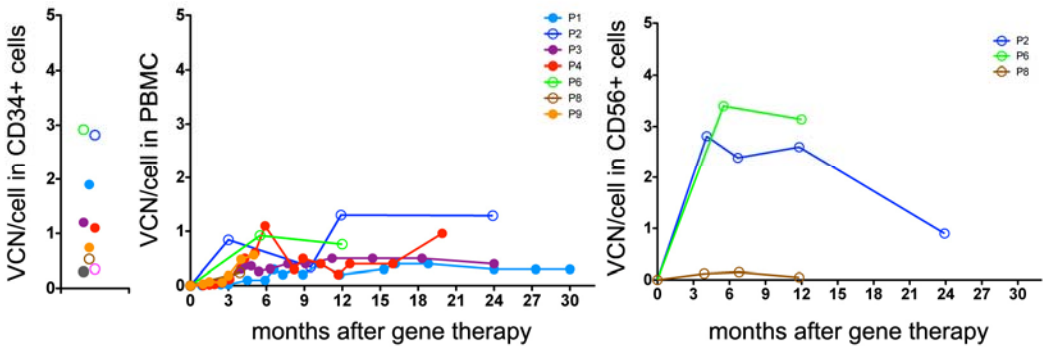
Supplemental Figure S1. Genetic map of MFG-γc and SIN-γc vectors.

MFG-γc vector backbone used in the original French trial is described in S. Hacein-Bey *et al.*,[8] and in the original British trial in Gaspar *et al*.[9] The MFG-LTR vector uses Moloney murine leukemia virus (MO-MLV) LTRs for transcription of the viral genome, and contains, an extended packaging (Ψ++) sequence, splice donor (SD) and splice acceptor (SA). In the French trial, the vector also contains the B2 mutation in the proline (P) primer binding site corresponding to a single G to A transition at position +160 of the MO-MLV sequence.

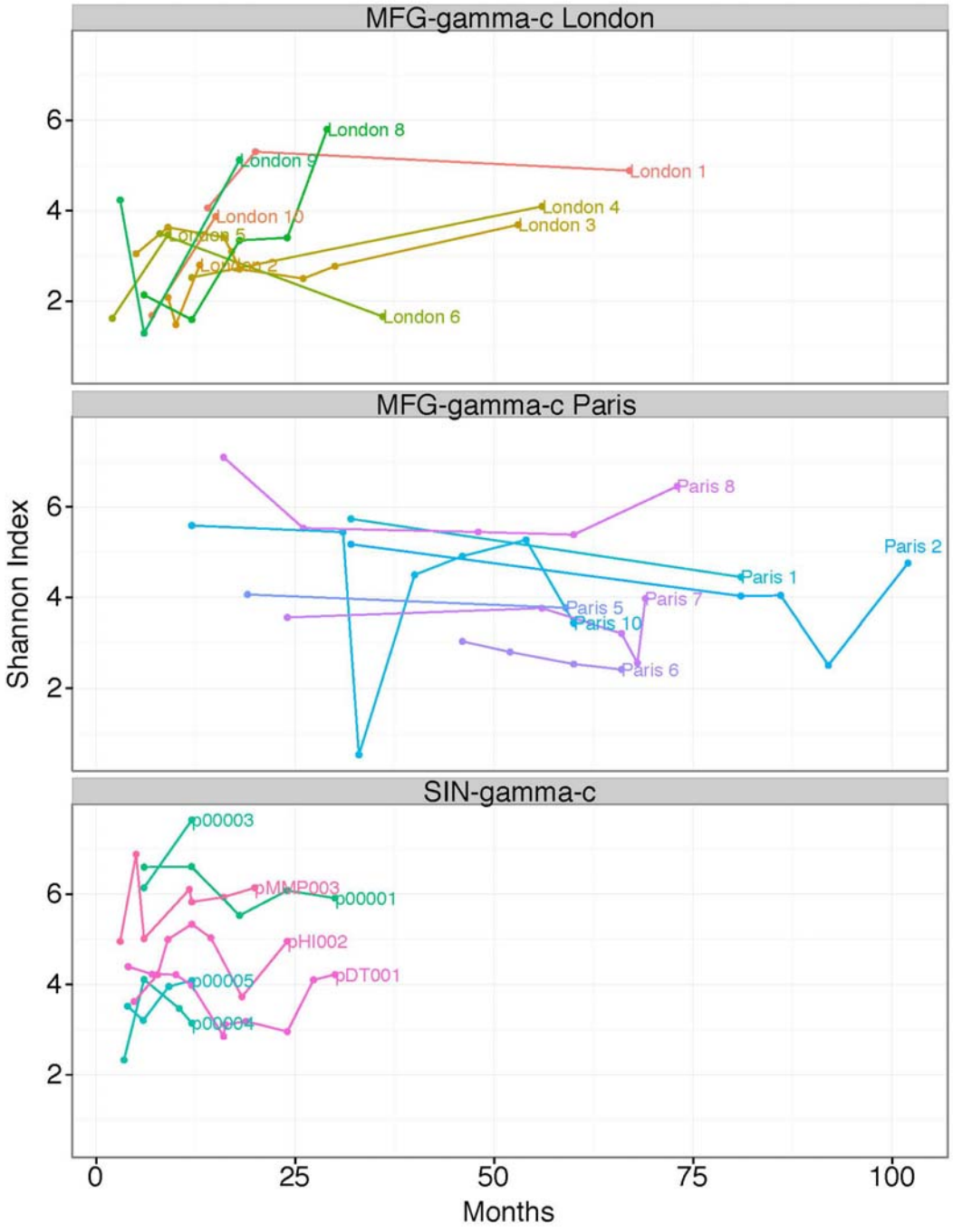
The SIN-γc vector backbone has been described.[10,11] Descriptive nomenclature as follows: SIN with RSV promoter, Leader 11, i.e. SIN vector with the RSV promoter fused to the R region (+1 relative to transcriptional start site). The LTR contains a SIN deletion (Δ) within the U3 region. The gag-free leader 11 (Hildinger et al. 1999, J Virol) is derived from Murine Embryonic Stem Cell Virus (MESV) and contains a glutamine (Q) primer binding site, a splice donor site and the packaging signal Ψ. Promoter from human elongation factor 1a short (EFS) contains 240bp elongation factor promoter without intronic sequences. IL2RG cDNA, flanked by *Age*I and *Sal*I sites. At the 5' terminus a partial Kozak sequence (GCCACCatg) was added. Post-Transcriptional Regulatory Element of Woodchuck Hepatitis Virus (WHV) for improvement of titer and gene expression. The PRE version used here (PRE*) does not contain sequences of X protein ORF. In addition, the largest ORF within the PRE (initiated by an ATG) is deleted.[12]



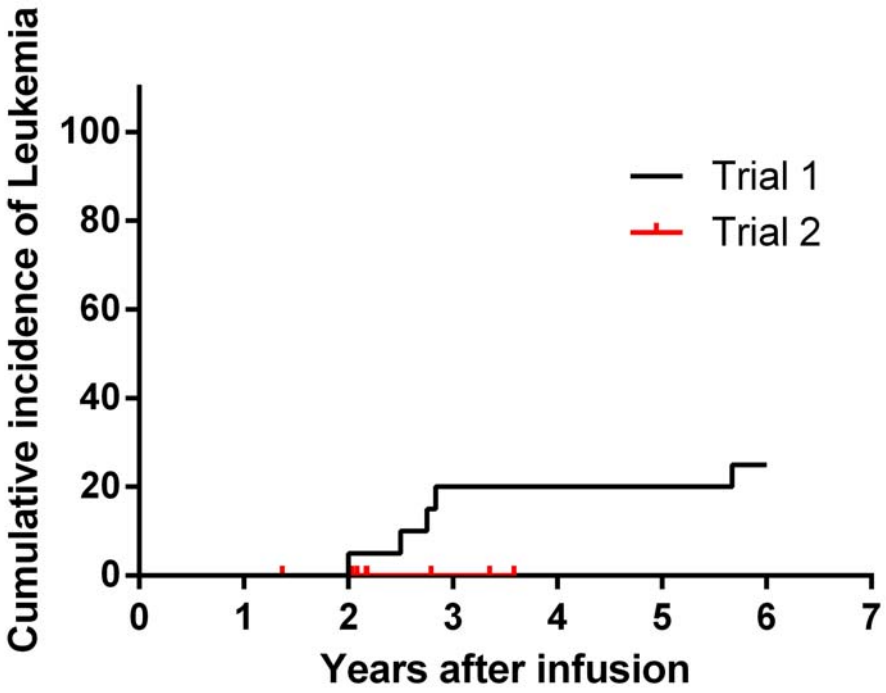
Supplemental Figure S2. Vector copy number (VCN) per cell in infused $CD34^+$ cell product, peripheral blood mononuclear cells (PBMC) and $CD3^-$ $CD56^+$ lymphocytes is shown.



Supplemental Figure S3. Diversity of cell clones inferred from integration site information. The diversity was inferred from the numbers of unique integration sites and the sequence counts for each. The trial analyzed is indicated at the top of the panels. The x-axis shows the time after cell infusion, the y-axis shows the Shannon index calculated for each patient/time point combination.



Supplemental Figure S4: Cumulative incidence of leukemia after infusion, for Trial 1 (n=20 from closed trials in London and Paris using MFG-γc vector) and Trial 2 (n=9 in the ongoing trial). On Trial 2, censored observations of current follow-up are shown with tick marks. On Trial 1, the 15 patients who have not gotten leukemia are censored beyond 6 years although tick marks are not shown, i.e., they have more than 6 years of follow-up.



Because the cumulative incidence of leukemia in our cohort is 0% at all timepoints, an upper limit of the incidence (i.e., risk) of leukemia is not estimable; therefore we performed a descriptive analysis. The 5.7-year estimated cumulative incidence of leukemia for the prior trials from London and Paris, where 5 leukemias (out of 20 patients) have been reported, was 25% with a 95% confidence interval of 12%-53%, using methods that would adjust for the competing risk of death. It is noteworthy that the lower limit of the 95% CI (12%) is greater than 0%. Further formal statistical comparison can be performed with full enrollment and longer follow-up.

Supplemental Table S1

Resolution of infections

Pt # Center	Age at GT (months)	Infections present at time of GT	Time of resolution (months)	Status
1	8.3	Disseminated BCGitis*	21.2	Infections resolved
		Pneumonitis	Prior to GT	
2	5.8	Oral ulcers, presumed viral*	2.2	Infections resolved
3	5.5	Disseminated BCGitis*	19.8	Infections resolved
		CMV infection*	0.2	
		EBV LPD*	0.2	
		RSV*	5.5	
4	6.8	Disseminated BCGitis*	16.2	Infections resolved
		Pneumocystis	Prior to GT	
5	9.0	Systemic severe adenovirus with hepatitis*	Did not resolve	Death secondary to respiratory failure and ongoing infection
6	10.5	Disseminated BCG*	24.2	Infections resolved
7	3.9	None	n/a	s/p cord blood transplant
8	8.2	Pneumocystis	Prior to GT	Infections resolved
9	8.0	Chronic diarrhea to rotavirus infection*	11.1	Infection resolved
		Pneumocystis	Prior to GT	

BCG: Bacille Calmette-Guérin; CMV: Cytomegalovirus; EBV: Epstein-Barr Virus; LPD: Lymphoproliferative disease

Supplementary References

1. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376-80.
2. Brady T, Roth SL, Malani N, et al. A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. Nucleic Acids Research 2011.
3. Berry C, Hannenhalli S, Leipzig J, Bushman FD. Selection of target sites for mobile DNA integration in the human genome. PLoS computational biology 2006;2:e157.
4. Ocwieja KE, Brady TL, Ronen K, et al. HIV Integration Targeting: A Pathway Involving Transportin-3 and the Nuclear Pore Protein RanBP2. PLoS pathogens 2011;7:e1001313.
5. Berry CC, Gillet NA, Melamed A, Gormley N, Bangham CR, Bushman FD. Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics 2012;28:755-62.
6. Berry CC, Ocwieja KE, Malani N, Bushman FD. Comparing DNA integration site clusters with scan statistics. Bioinformatics 2014.
7. Sadelain M, Papapetrou EP, Bushman FD. Safe harbours for the integration of new DNA in the human genome. Nature reviews Cancer 2011.
8. Hacein-Bey S, Cavazzana-Calvo M, Le Deist F, et al. gamma-c gene transfer into SCID X1 patients' B-cell lines restores normal high-affinity interleukin-2 receptor expression and function. In: Blood; 1996:3108-16.
9. Gaspar HB, Parsley KL, Howe S, et al. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. Lancet 2004;364:2181-7.
10. Schambach A, Mueller D, Galla M, et al. Overcoming promoter competition in packaging cells improves production of self-inactivating retroviral vectors. In: Gene Ther; 2006:1524-33.
11. Thornhill SI, Schambach A, Howe SJ, et al. Self-inactivating gammaretroviral vectors for gene therapy of X-linked severe combined immunodeficiency. Mol Ther 2008;16:590-8.
12. Schambach A, Bohne J, Baum C, et al. Woodchuck hepatitis virus post-transcriptional regulatory element deleted from X protein and promoter sequences enhances retroviral vector titer and expression. Gene therapy 2006;13:641-5.

Supplemental Report
Clustering of Integration Sites

June 15, 2014

Contents

Abstract

This report considers whether the different vectors used in two different trials resulted in different patterns of insertion of integration sites. As a secondary goal, the differences between patients seen in Britain and those seen in France during the first trial are considered. It updates a similar study from October 2012.

It is determined that individual patients in each trial and in Britain and in France differ from their peers in the terms of insertion site clustering. That is, a patient is more likely to have two or more insertion sites in nearby locations than would be predicted if all sites derived from a single vector were located and recovered independently (in the stochastic sense). This has implications for the design and analysis of trials in which insertion site locations are monitored.

Given patient level differentials, significance tests and other procedures for assessing error need to allow for patient level differences rather than treating counts of sites as independent entities. Here this is accomplished by using permutation procedures in which the trial labels are permuted among patients, so that every patient's sites are assigned to the same trial in any permutation.

Clusters or 'clumps' were identified as differing between the two trials. In the British to French patients in the first trial, clumps were identified, but the false discovery rate was too high to claim any as discoveries.

Comparison of site locations to the locations of genes in several gene lists suggested an association as did consideration of whether clumps from one trial were differentially found.

1 Data Basics

The data for this study were reduced to a set of 13373 sites unique in each patient in the first trial and 29100 in the second trial. The UCSD `hg18` freeze of the human genome was used[Lander et al., 2001].

Some integration sites were situated very close to others in the same patient and orientation (but not the reverse orientation), which is highly unlikely. As it seemed likely that those sites were mismapped clones of sites

already counted and thus could cause a spurious inference of clustering, they were removed. There were 126 removed from trial 1. There were 253 removed from trial 2.

The insertion sites were ordered by position on each chromosome, where 'position' refers to the site of attack farther from *pter*. When there were ties (i.e. two patients had insertions in the same location or one had a second insertion at the same location but in opposite orientation) the order was randomized. Adjacent pairs of sites were compared as follows: sites were labelled as

F patient treated in France during the first trial

L patient treated in London during the first trial

2 patient in the second trial

The tally of all adjacent pairs is presented in the following table along with the expected number of such pairs assuming no association among pairs. As can be seen there were more pairs of F follows F, L follows L and 2 follows 2 than expected. Thus, there is a suggestion that the spatial preferences of the vectors used were different.

```
##     post        F               L               2
##          Observed Fitted Observed Fitted Observed Fitted
## pre
## F            2799   2267      859    806     6106   6691
## L             829    805      454    286     2185   2376
## 2            6139   6694     2158   2379    20529  19753
```

2 Differential Clustering or Clumps

The approach taken here is is described in detail elsewhere [Berry et al., 2014]; it searches for regions defined by contiguous insertion sites in which the first or the second SCID trial showed up more often than expected. In addition, sites are ignored that are spread over a region occupying a wider interval as they are likely to be of less interest regardless of the trial in which they were seen. The usual caution about genomewide searches and the large number of tests - and possible false discoveries - applies here, so some measures must be taken to control and assess the error rates.

It is useful to model the process governing differentials in the appearance of sites from different trials using the Poisson Clumping Heuristic

[Aldous, 2010], which posits that a spatial process with local (but no long range) dependence can be treated as a Poisson proccess in which the *clumps* (or local collections) are treated as the events. Here, windows covering k adjacent insertion sites are considered, and the number from each trial in each window is tallied. Windows in which the number from one trial is remarkably large are marked. These marks tend to be locally correlated — if the k adjacent sites in one window are mostly from one trial, the next window will cover k-1 of the same sites and likely also have a large number of sites from that trial. Grouping adjacent windows showing a differential into *clumps* and then treating the clumps as the units that are discovered obviates the local dependence of the process by which windows are marked. An expected discovery rate, λ, is determined from a permutation null distribution. Following [Siegmund et al., 2011], the false discovery rate is taken as

$$\hat{\text{FDR}} = \frac{\lambda}{1+R}$$

where R is the actual number of discoveries claimed. The numbers of insertion sites counted in a window were k = 15, 16, ..., 75, and for each window size the median number of bases it spanned was computed. Only windows covering fewer bases than the median were screened. The value of λ was determined by permuting the trial labels identities between patients and counting the number of clumps discovered; the mean of that count over 1000 replicates is taken as the estimate of λ. Initially, a target of 5 False Discoveries per window is set, but assuming that insertions are independent, identically distributed regardless of the patient who contributed them. However, that assumption is not reasonble — the number discovered under permutation was more than twice that much, so the initial results were pruned to only include those passing below a target of 0.025 False Discoveries which corresponded to a estimated false discovery rate of around five percent. This procedure was also performed for the comparison of **B** versus **F** patient in the first trial using 1 False Discovery as the target.

This resulted in 21 clumps being discovered at an estimated False Discovery Rate of 0.041. The following table shows the locations of the clumps (per the `hg18` freeze), the number of sites from each trial in the clump, the *depth* or maximum number of different window widths covering the clump, the natural logarithm of the odds ratio for each clump (which would be the same as the logarithm of the odds if there were equal numbers of sites from the two trials), and the FDR for surpassing the target FDs achieved for that clump. The results are in order - lowest FDR first. The results for the **B** ver-

sus **F** patients are not given in detail; there were 6 clumps discovered at an estimated FDR of 0.348, which is too large to lead to confident exploration of the discovered clumps.

```
##  Chromosome     start   width SCID1 SCID2 depth log.OR     FDR
##        chr3 170335642  628631    45     6    61  -2.72 0.00000
##       chr12   4122059  249349    50    10    61  -2.35 0.00133
##       chr11  33849301  234851    56    24    61  -1.61 0.01380
##       chr17  43727056  541728    42    14    61  -1.85 0.01380
##        chr7  26405521  632529    40    13    61  -1.88 0.01533
##       chr12     91299  636177    37    16    61  -1.60 0.02443
##       chr11  34403251  653983     1    33     4   2.33 0.03220
##       chr12  64503484  101771    19     3    55  -2.50 0.03220
##       chr16  23796898   72675    26     7    60  -2.04 0.03220
##       chr14  34825527   81420    29    10    60  -1.81 0.03436
##        chr1 232725649  511081    84    65    61  -1.03 0.03700
##        chr1 148604772  248140    22     6    37  -2.02 0.03700
##       chr19  55998313 3777743     7    90    52   1.71 0.03700
##        chr1  54615965  837829    25     8    61  -1.88 0.03700
##        chr6  15375728   97206    17     2    61  -2.72 0.03700
##        chr5 139007263   61363    33    16    54  -1.49 0.03700
##       chr20  29619021  472656    32    15    56  -1.52 0.03700
##       chr17  72840348  240060    38    22    49  -1.31 0.03926
##       chr14  76535994   34427     9     0    59  -3.72 0.03946
##       chr15  73108148  151760    28    12    51  -1.60 0.03946
##       chr14  49443909   93728    25     9    57  -1.77 0.03946
```

3 Sample Scheduling Differences and Clumping

The data collection days for the two trials differed. The *SCID2* data have so far been collected on or before day 915 and most *SCID1* integration sites were first seen after that time. If integration sites of low abundance in a genomic locale experienced proliferation later on, it is possible that apparent clumps composed of sites from the *SCID1* trial would disappear when later *SCID2* data are acquired. The graph shown in Figure 1 depicts the distributions of first and last observations of integration sites according to the trial in which they were found and whether they were in clumps discovered at moderate or low targets for expected False Discoveries. As can be seen the distributions within each trial are very similar. If clumps

only occurred after late outgrowth of low abundance clones one would expect that the distributions for clones would skew towards later detection, but this skewing is not apparent.

It is possible that there are both apparent clumps that are artifacts of different data collection schedules and some clumps that would have been seen with identical schedules. So a clump-by-clump inspection is necessary to rule this out. The next table shows the first and third quartiles for the first day on which a site was detected in *SCID1* and the last day for *SCID2*. A particularly eyecatching clump is on chromosome 11 starting at position 34403251. However, this particular clump favors integration of *SCID2* sites and there was only one *SCID1* site in the clump. So, the data collection schedule could not have induced this result. `NA` values seen in the table arise when there are no integration sites for one of the trials.

```
##  Chromo     start log.OR SCID1.25 SCID1.75 SCID2.25 SCID2.75
##    chr3 170335642   -2.7      488     1464       91      640
##   chr12   4122059   -2.3      488      976        0      206
##   chr11  33849301   -1.6      488      976        0      396
##   chr17  43727056   -1.9      503      976      152      915
##    chr7  26405521   -1.9      488     1342      183      366
##   chr12     91299   -1.6      488     1662        0      366
##   chr11  34403251    2.3     1586     1586        0      488
##   chr12  64503484   -2.5      557      930      530      915
##   chr16  23796898   -2.0      976     1708      259      486
##   chr14  34825527   -1.8      732     1876        0      457
##    chr1 232725649   -1.0      488     1464        0      488
##    chr1 148604772   -2.0      366     1647      518      701
##   chr19  55998313    1.7      366     1403        0      366
##    chr1  54615965   -1.9      549     1830      366      607
##    chr6  15375728   -2.7      488     1464      229      320
##    chr5 139007263   -1.5      778     1830       69      366
##   chr20  29619021   -1.5      610     2044        0      365
##   chr17  72840348   -1.3      580     1708      152      366
##   chr14  76535994   -3.7      488     1159       NA       NA
##   chr15  73108148   -1.6      488     1807        0      365
##   chr14  49443909   -1.8      488     1342      152      488
```

A more direct method of addressing the issue of differences in schedules is to truncate the data for the *SCID1* trial so that it better matches the collection schedule for the *SCID2* trial. Here the data are truncated at day

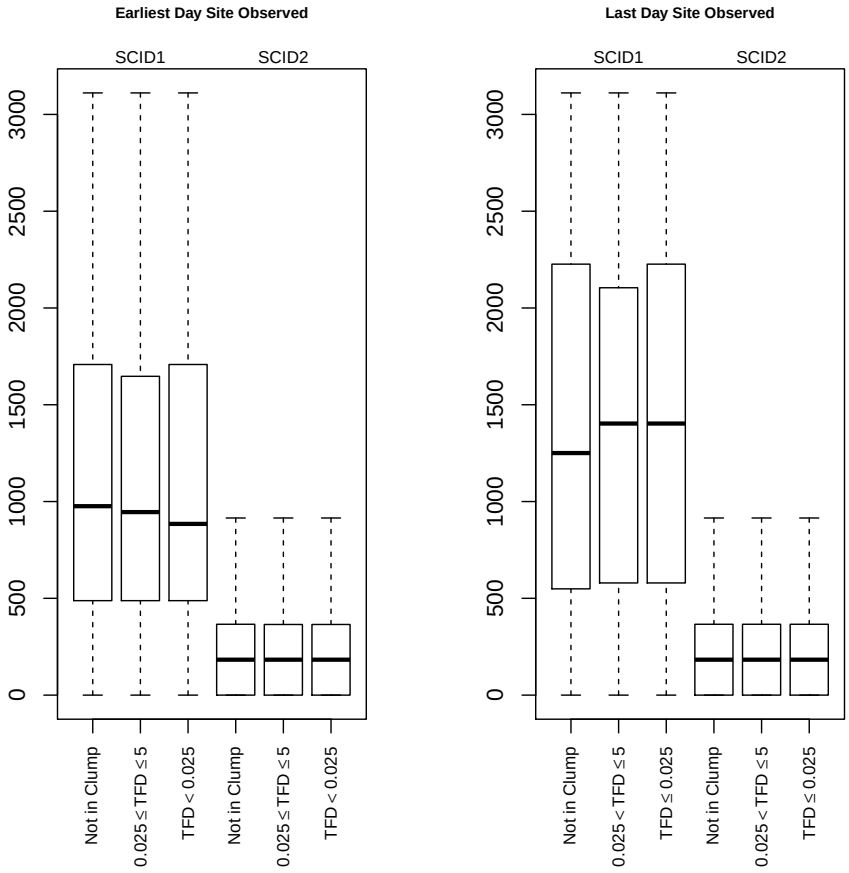


Figure 1: Observation Times for Integration Sites: Boxes show the quartiles of the distribution, the median in each box, and the extremes are shown as whiskers (when they extend beyond the quartiles). Groupings are formed according to trial and according to whether the sites fell into clumps and if so whether the target for False Discoveries (TFD) was less than 0.025.

920 and the clump discovery is carried out using the same parameters. This truncation does not balance the collection schedules, but does assure that differences that occur more than 2.5 years post-transplant will not sway the results. The table below shows the 16 clumps discovered that overlap the collection of 21 clumps displayed above. The first 5 clumps from earlier (with the smallest FDRs) are all seen here. Of course, culling the data in this way reduces the number of integration sites available for analysis in the *SCID1* trial. There are 13247 *SCID1* sites in the complete data, but only 6072 sites seen before day 920. Obviously, this reduces the power to detect clumping and might be responsible for the failure to discover several clumps seen in the complete data.

```
##  Chromosome     start   width SCID1 SCID2 Target Min
##        chr1  54610189  843605    13     9   6.66e-01
##        chr1 148801899   51891    13     5   2.23e-03
##        chr1 232725675  511050    47    65   3.65e-03
##       chr11  33849296  234856    38    24   1.09e-09
##       chr12   3913563  457845    35    15   3.29e-11
##       chr12  64238360  587130    19    14   5.02e-04
##       chr14  49400145  137492    16    12   1.32e-01
##       chr14  76462472  115481    19    19   2.34e-01
##       chr15  72985467 2090417    24    31   3.06e-01
##       chr17  43676435  353639    22    15   1.68e-04
##       chr19  53829165 2613299     2    69   3.59e+00
##       chr19  57723157 6056204     2   102   1.15e+00
##       chr20  29619016  472656    15    15   2.19e+00
##        chr3 170337883 1218510    24    10   4.14e-08
##        chr6  15375728  117254    10     5   1.48e+00
##        chr7  26506343  170688    16     7   2.45e-03
```

The current data do not positively rule out the possibility that later followup of the *SCID2* patients will result in the diminution of clumps seen in the curent data, but there is no obvious signal in these data that the difference in the collection schedules has artifactually produced clumps.

4 Delicate Regions

Certain regions of the genome may be deemed to be *delicate* in the sense that an insertion in them might precipitate an adverse event. Here several types

of regions are considered and for each the relative frequency of insertions from the two trials is compared. The regions studied are:

Lymphoid Cancer Associated Genes The bodies of genes associated with lymphoid cancers (abbreviated LymCaBody)

Lymphoid Cancer Associated Gene Regions the 50 kilobase regions preceding or succeeding lymphoid cancer associated genes (abbreviated LymCaEdge)

All Cancer Associated Genes The bodies of genes associated with cancer (abbreviated AllCaBody)

All Cancer Associated Gene Regions the 50 kilobase regions preceding or succeeding all cancer associated genes (abbreviated AllCaEdge)

In this table, the '+' sign indicates that the clump covered one (or more) of the regions in question. The FDR criterion was relaxed to include 45 clumps discovered at an estimated False Discovery Rate of 0.104. The number of clumps covering regions was compared to the distribution formed by permutation of patient labels across trials; the p-values are shown below the table. The Holm adjusted two-sided p-values are shown, and 2 of the 4 results are statistically significant at $p < 0.05$.

```
##  LymCaBody LymCaEdge AllCaBody AllCaEdge log.OR
##          -         -         +         + -2.724
##          +         +         +         + -2.349
##          +         +         +         + -1.614
##          -         -         +         + -1.854
##          -         -         -         - -1.877
##          -         -         +         + -1.599
##          -         -         -         -  2.328
##          -         -         +         + -2.496
##          -         -         -         - -2.040
##          -         -         -         + -1.811
##          -         -         +         + -1.033
##          -         -         +         + -2.020
##          -         -         +         +  1.712
##          -         -         +         + -1.877
##          -         -         -         - -2.724
##          -         -         -         + -1.486
```

```
##          -         -         +         + -1.519
##          -         -         +         + -1.315
##          -         -         -         - -3.723
##          -         -         -         - -1.602
##          -         -         -         + -1.766
##          -         -         +         + -1.156
##          -         -         -         - -2.329
##          -         -         +         +  1.967
##          -         -         +         +  1.369
##          -         -         -         -  3.484
##          -         -         +         + -2.078
##          -         -         -         - -1.584
##          -         -         +         + -1.016
##          -         -         +         + -0.976
##          +         +         +         + -2.536
##          -         -         +         +  2.078
##          -         -         +         +  1.395
##          -         -         +         - -1.584
##          -         -         +         + -1.640
##          -         -         +         - -2.266
##          -         -         -         - -1.149
##          -         -         -         - -1.300
##          -         -         -         -  3.333
##          -         -         -         - -0.958
##          -         -         -         - -2.200
##          -         -         +         +  1.226
##          -         +         -         + -1.176
##          -         -         +         + -1.133
##          -         -         -         - -2.200
```

```
## LymCaBody LymCaEdge AllCaBody AllCaEdge
##     0.150     0.068     0.036     0.024
```

The following tables show how many insertions fell in each of the types of regions according to trial. The Holm adjusted two-sided p-values are

```
## LymCaBody LymCaEdge AllCaBody AllCaEdge
##     0.006     0.004     0.012     0.384
```

```
## $LymCaBody
##   FirstSCID sinSCID FirstSCID sinSCID
## -     13124   28678    0.9907  0.9941
## +       123     169    0.0093  0.0059
##
## $LymCaEdge
##   FirstSCID sinSCID FirstSCID sinSCID
## -     13058   28677    0.9857  0.9941
## +       189     170    0.0143  0.0059
##
## $AllCaBody
##   FirstSCID sinSCID FirstSCID sinSCID
## -     11384   25232    0.8594  0.8747
## +      1863    3615    0.1406  0.1253
##
## $AllCaEdge
##   FirstSCID sinSCID FirstSCID sinSCID
## -     10961   24073    0.8274  0.8345
## +      2286    4774    0.1726  0.1655
```

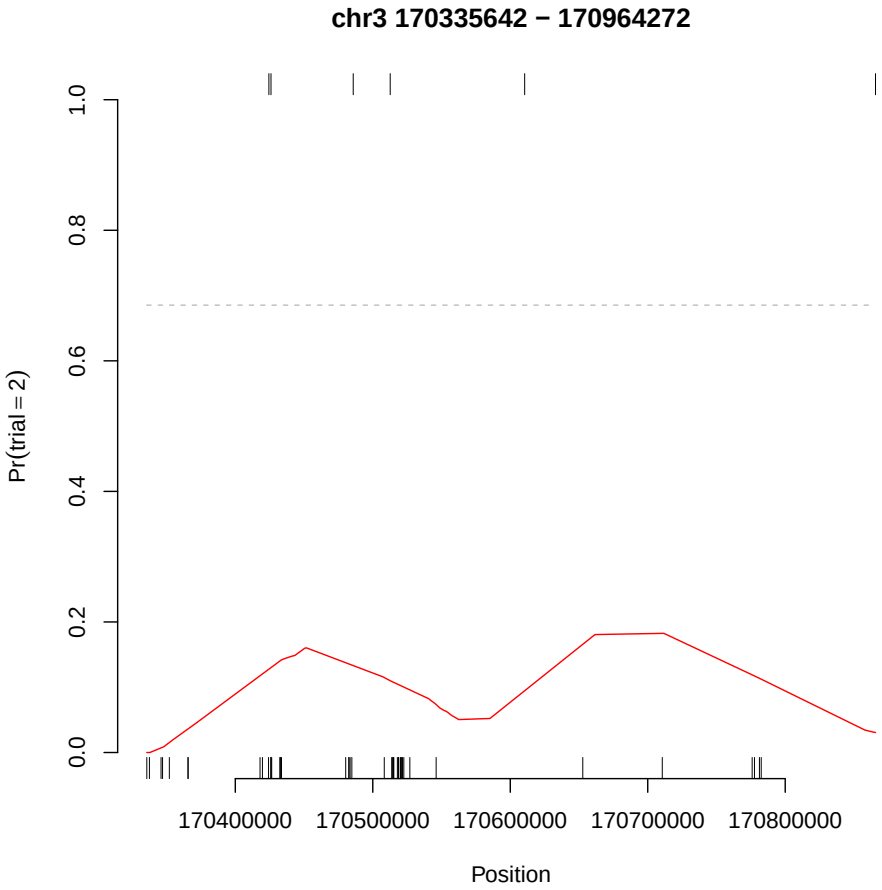
5 Detailed Results

The details of each clump discovered at an FDR of 0.041 or less are illustrated in the following pages of graphs. Each figure shows

- the locations of sites from trial 1 as tickmarks on the bottom axis
- the locations of sites from trial 2 as tickmarks on the top axis
- the overall proportion of sites from trial 2 as a line across the figure
- a rough estimate of the proportion of sites from trial 2 in the local region (as a loess fit constrained to the unit interval) as a red curve. The tickmarks contain the most information, but can be hard to perceive. The fitted curve is intended only as an aid to rapid viewing.
- summary statistics for the clump in the gray panel at the top

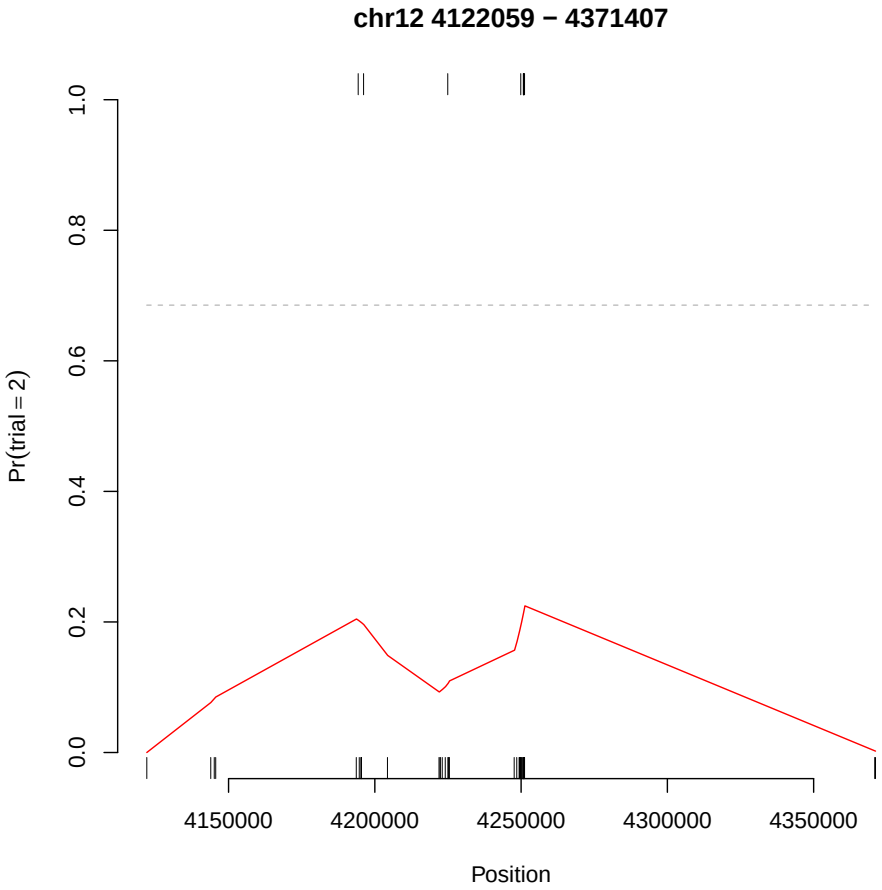
5.1 Clump chr3:170335642-170964272

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr3	170335642	628631	45	6	61	-2.72	0



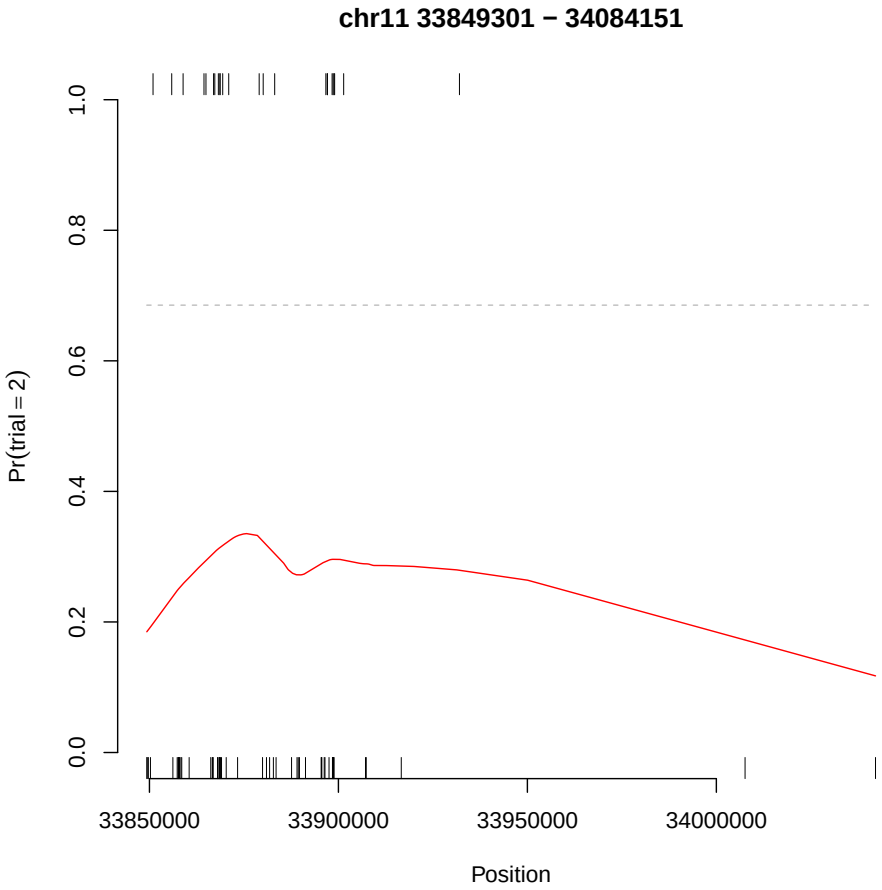
5.2 Clump chr12:4122059-4371407

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr12	4122059	249349	50	10	61	-2.35	0.00133



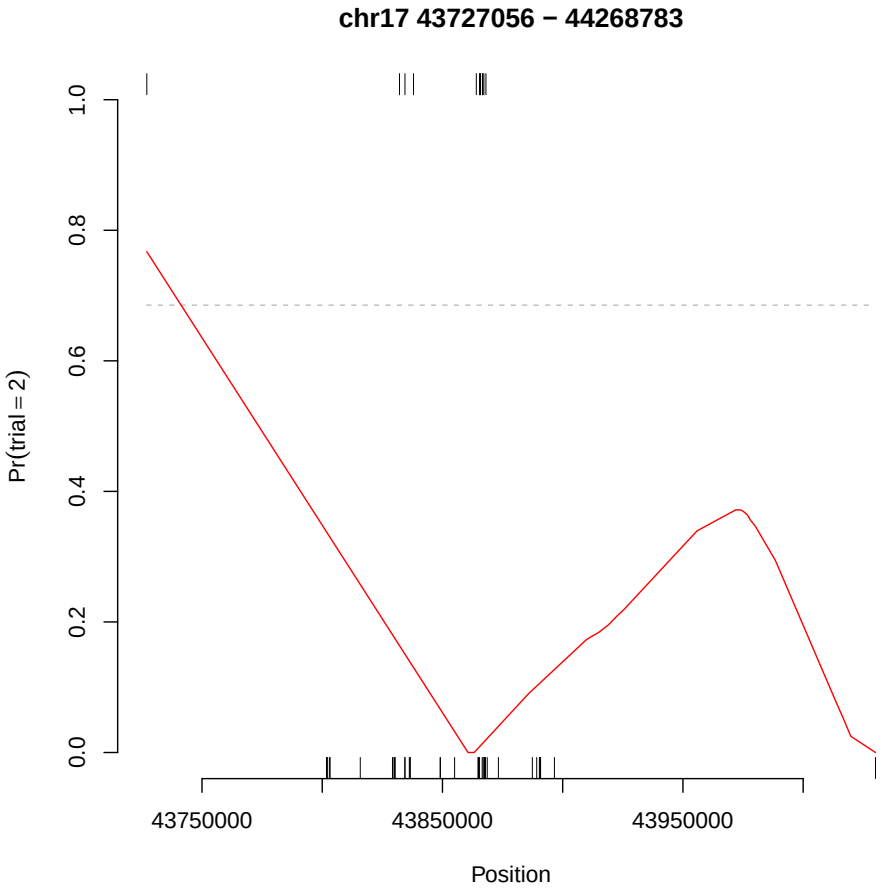
5.3 Clump chr11:33849301-34084151

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr11	33849301	234851	56	24	61	-1.61	0.0138



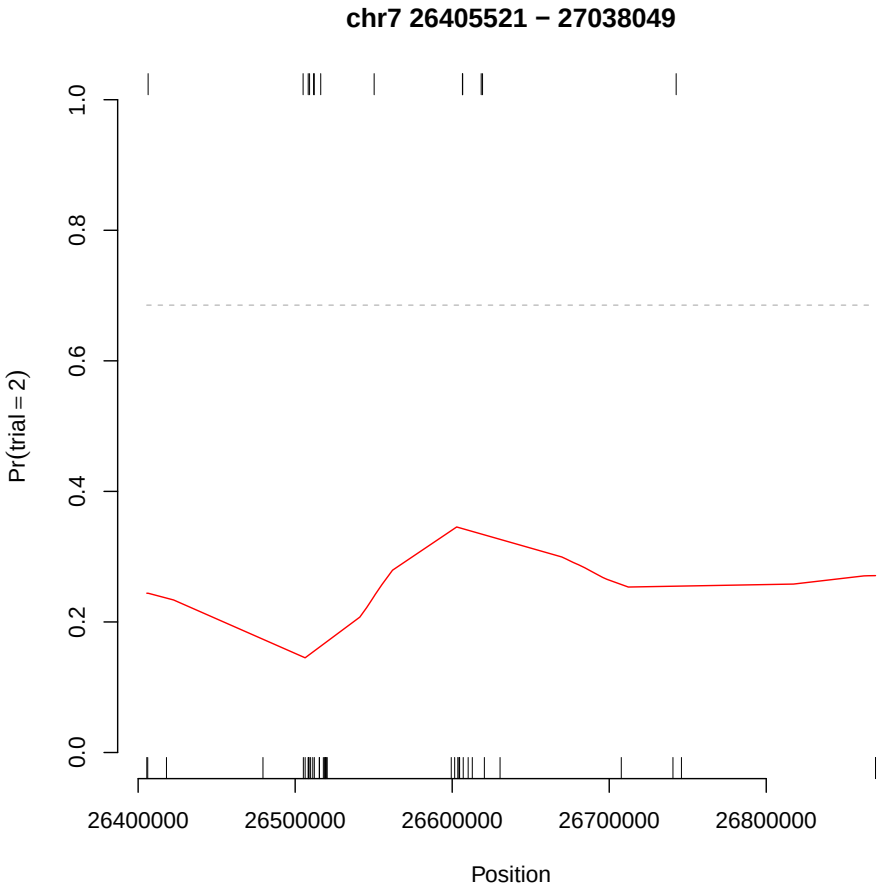
5.4 Clump chr17:43727056-44268783

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr17	43727056	541728	42	14	61	-1.85	0.0138



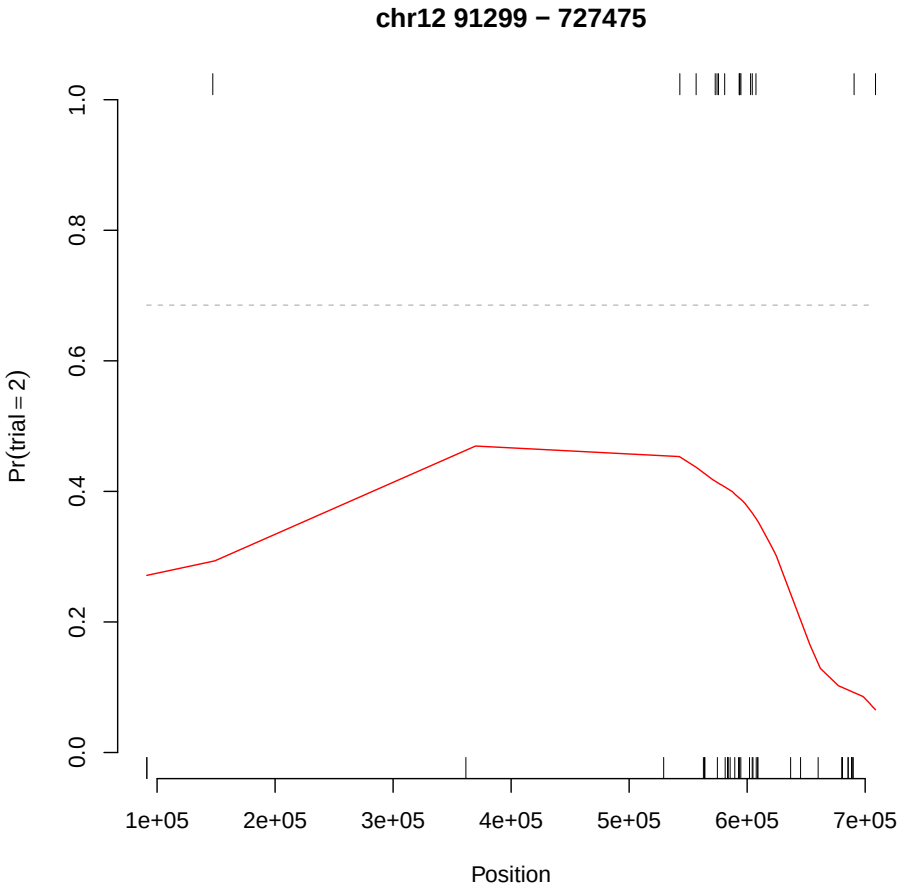
5.5 Clump chr7:26405521-27038049

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr7	26405521	632529	40	13	61	-1.88	0.0153



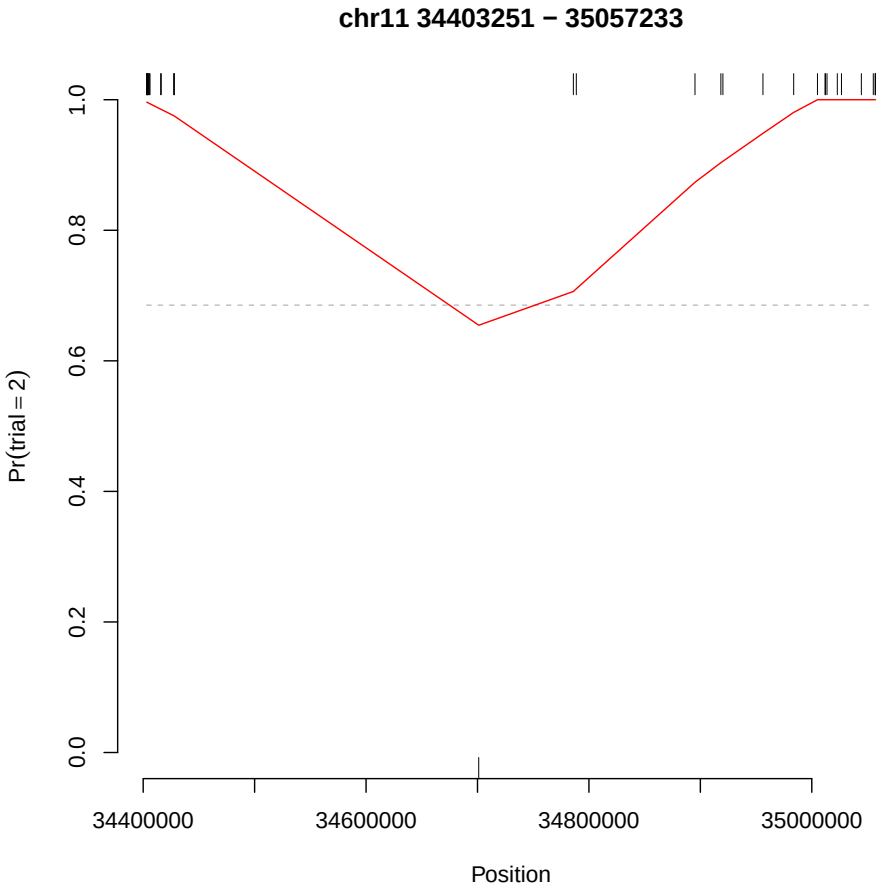
5.6 Clump chr12:91299-727475

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr12	91299	636177	37	16	61	-1.6	0.0244



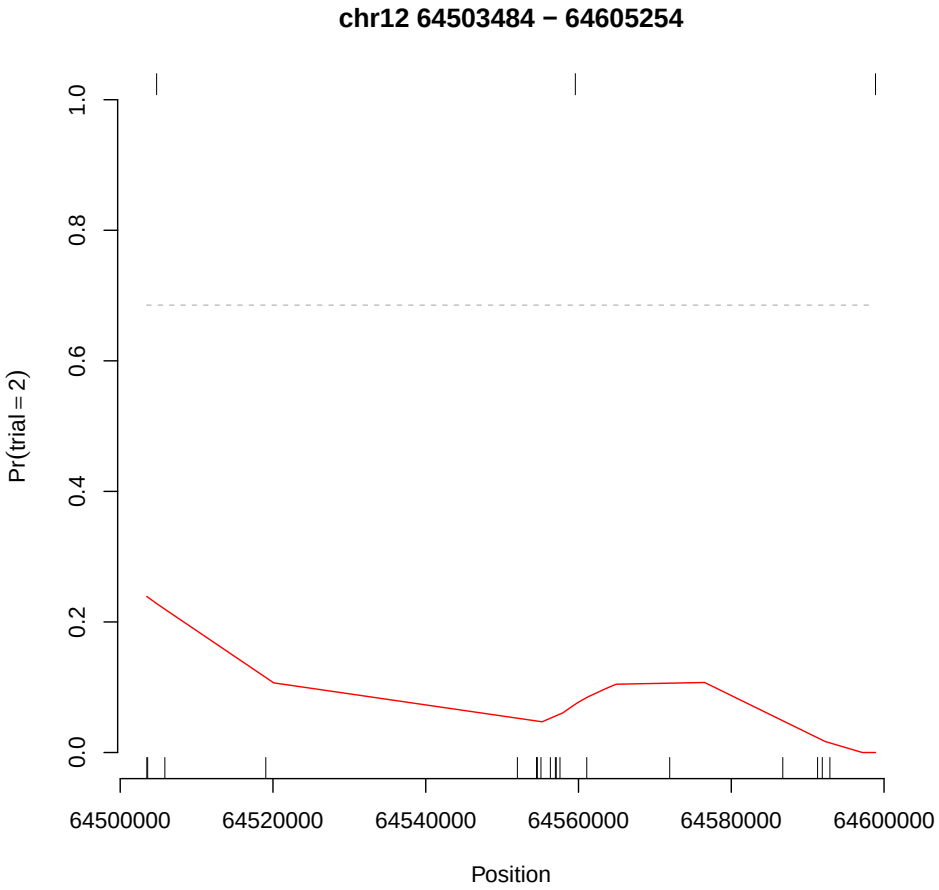
5.7 Clump chr11:34403251-35057233

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr11	34403251	653983	1	33	4	2.33	0.0322



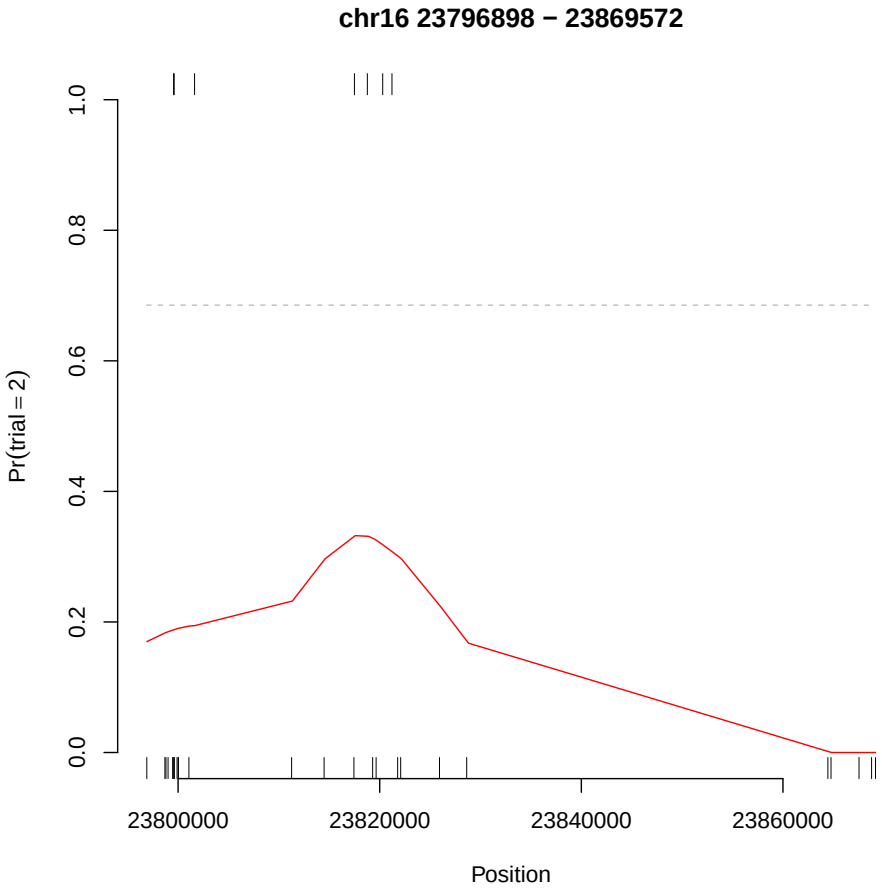
5.8 Clump chr12:64503484-64605254

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr12	64503484	101771	19	3	55	-2.5	0.0322



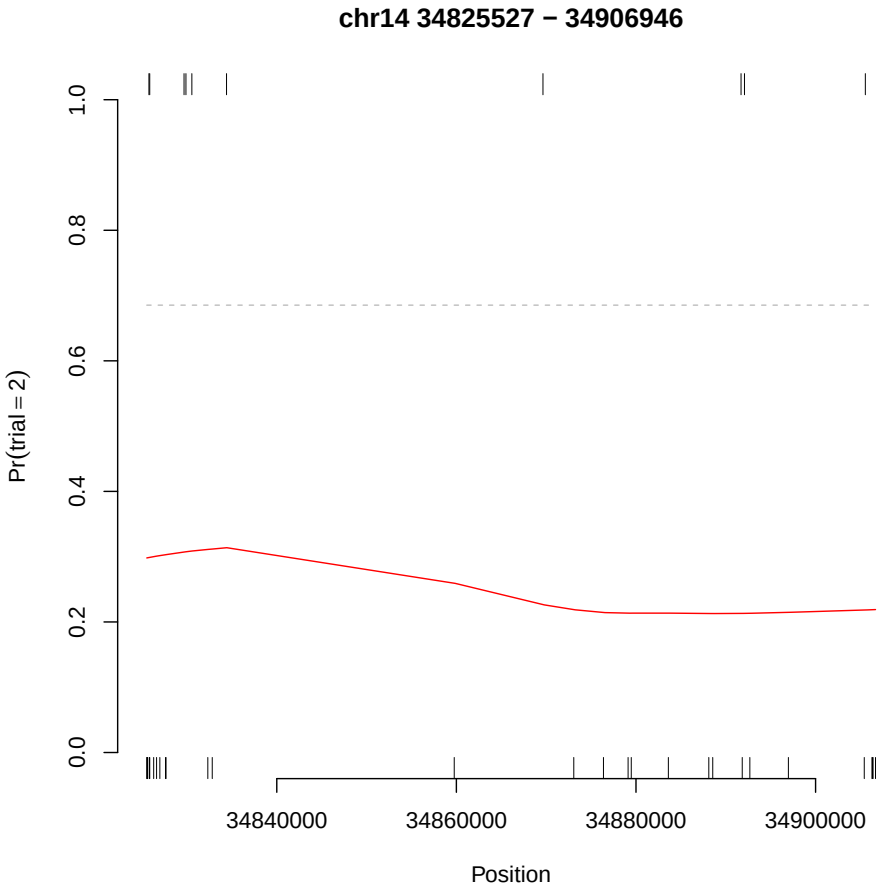
5.9 Clump chr16:23796898-23869572

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr16	23796898	72675	26	7	60	-2.04	0.0322



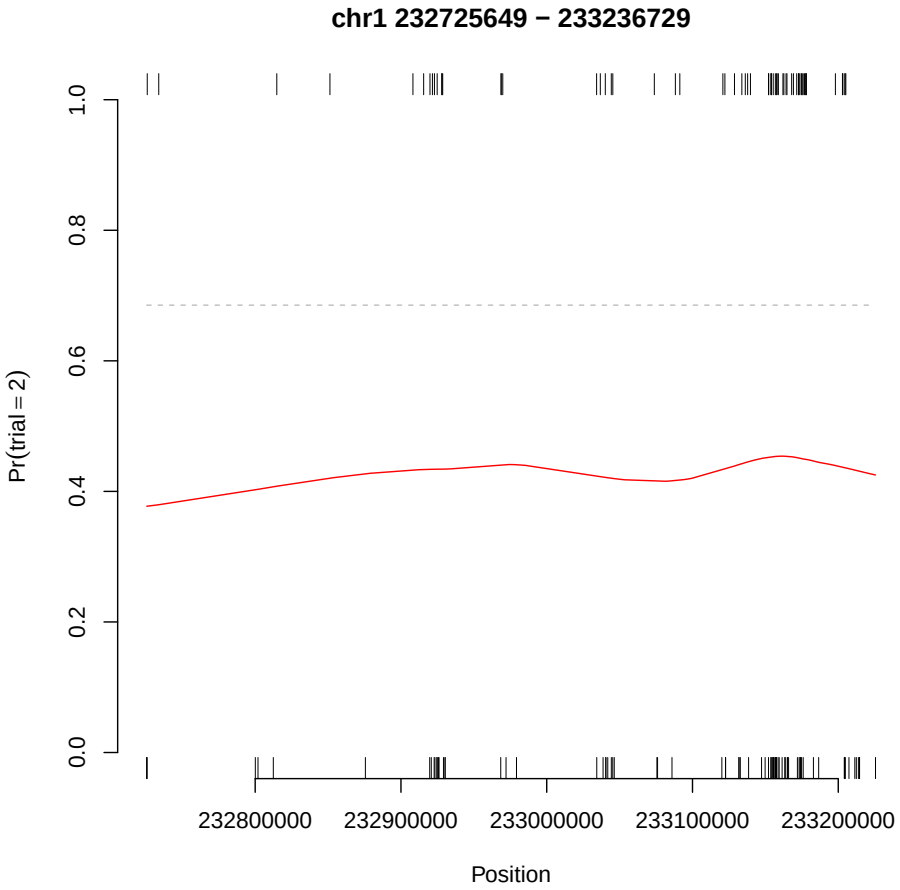
5.10 Clump chr14:34825527-34906946

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr14	34825527	81420	29	10	60	-1.81	0.0344



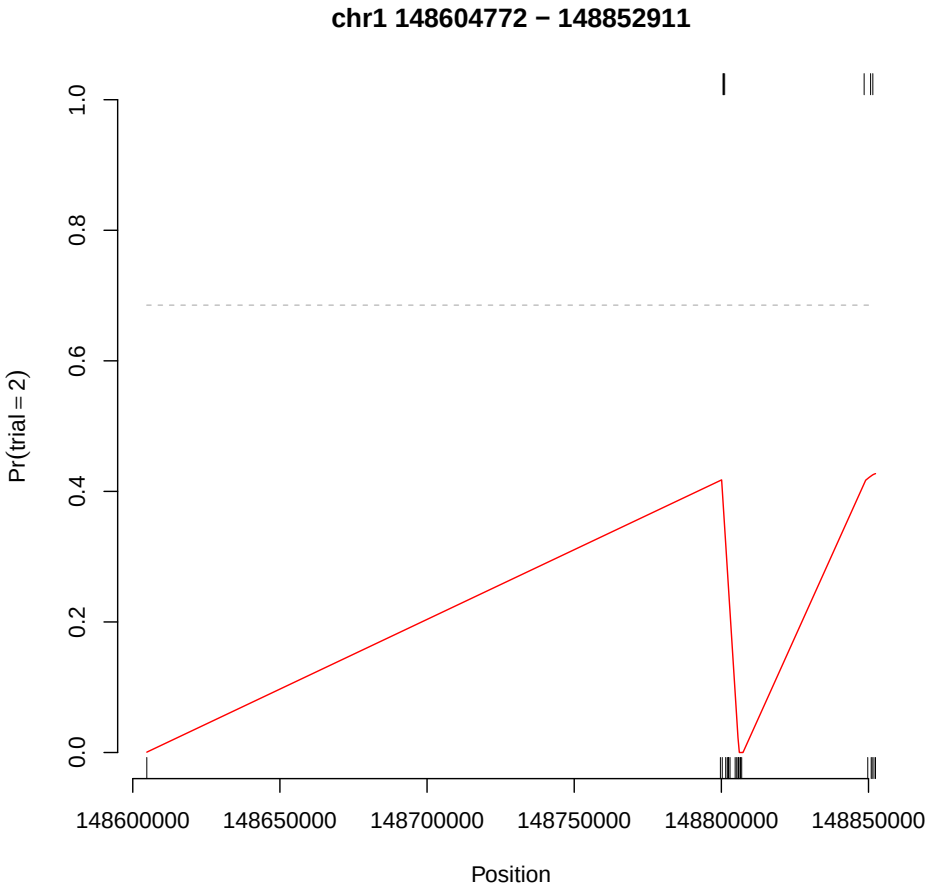
5.11 Clump chr1:232725649-233236729

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr1	232725649	511081	84	65	61	-1.03	0.037



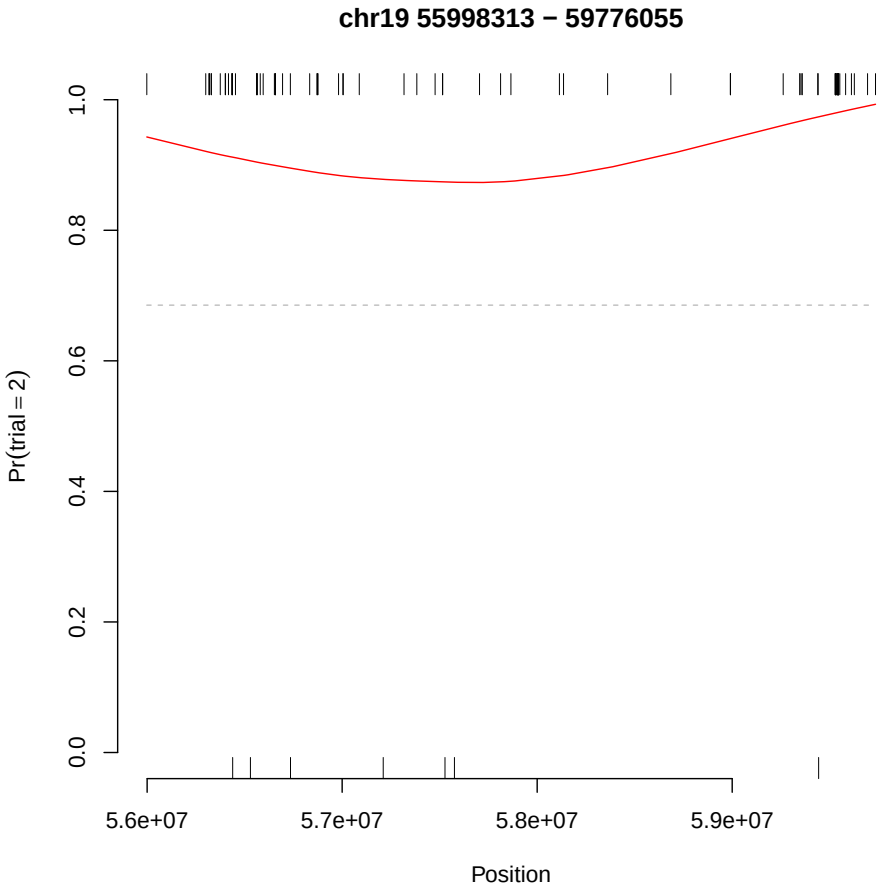
5.12 Clump chr1:148604772-148852911

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr1	148604772	248140	22	6	37	-2.02	0.037



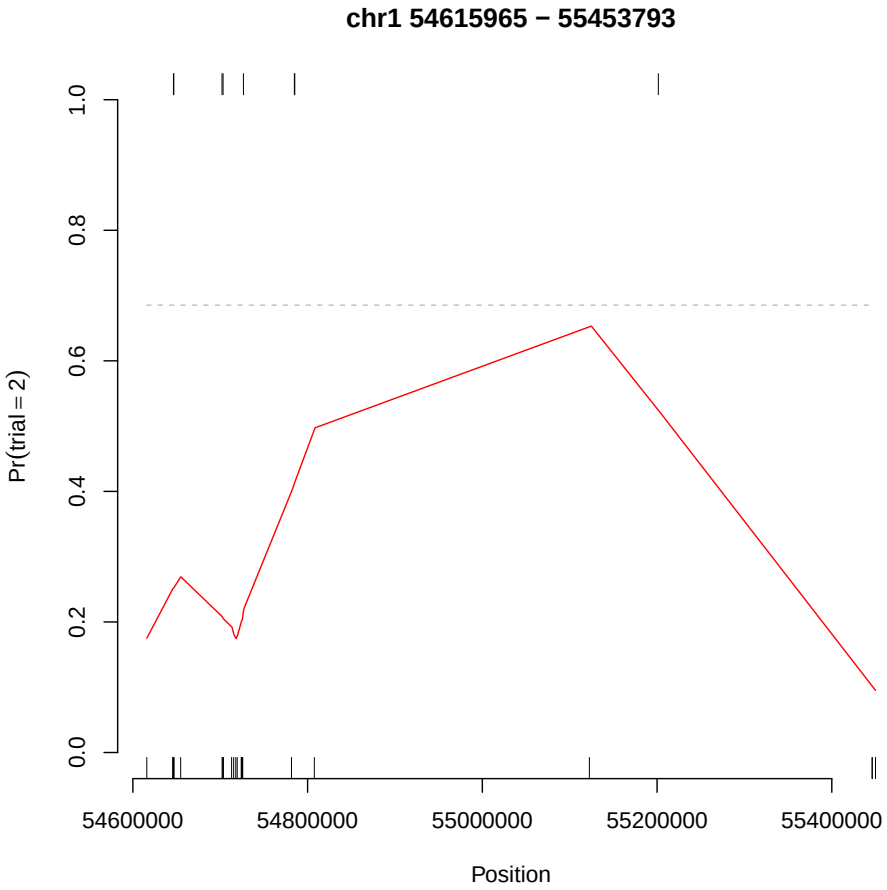
5.13 Clump chr19:55998313-59776055

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr19	55998313	3777743	7	90	52	1.71	0.037



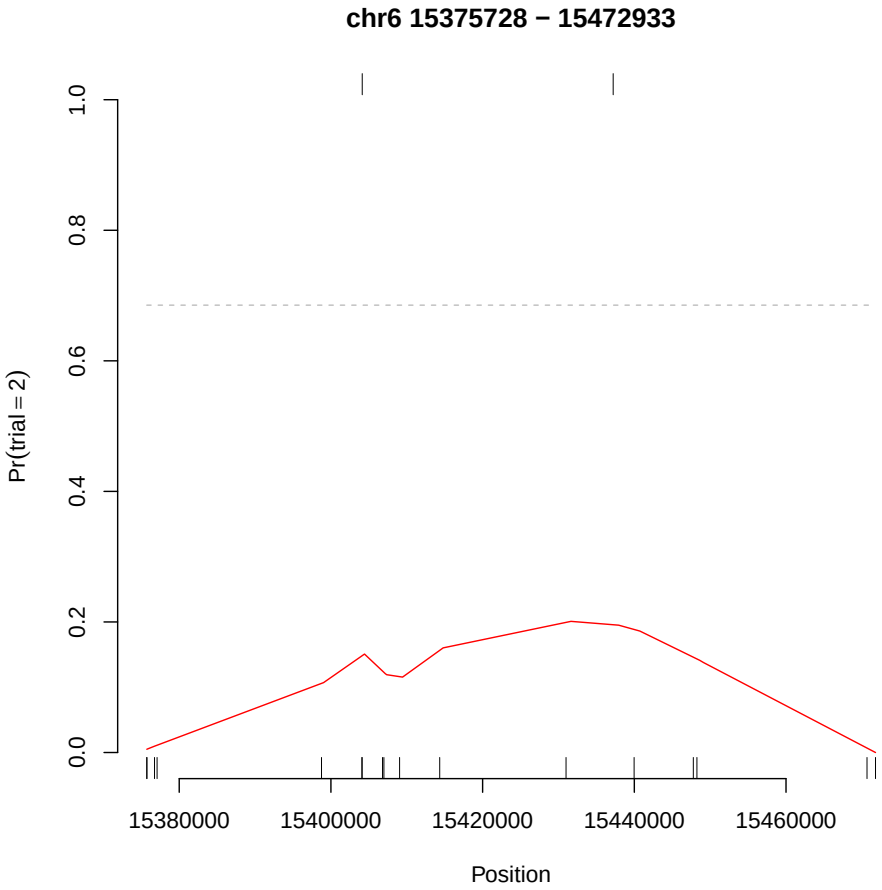
5.14 Clump chr1:54615965-55453793

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr1	54615965	837829	25	8	61	-1.88	0.037



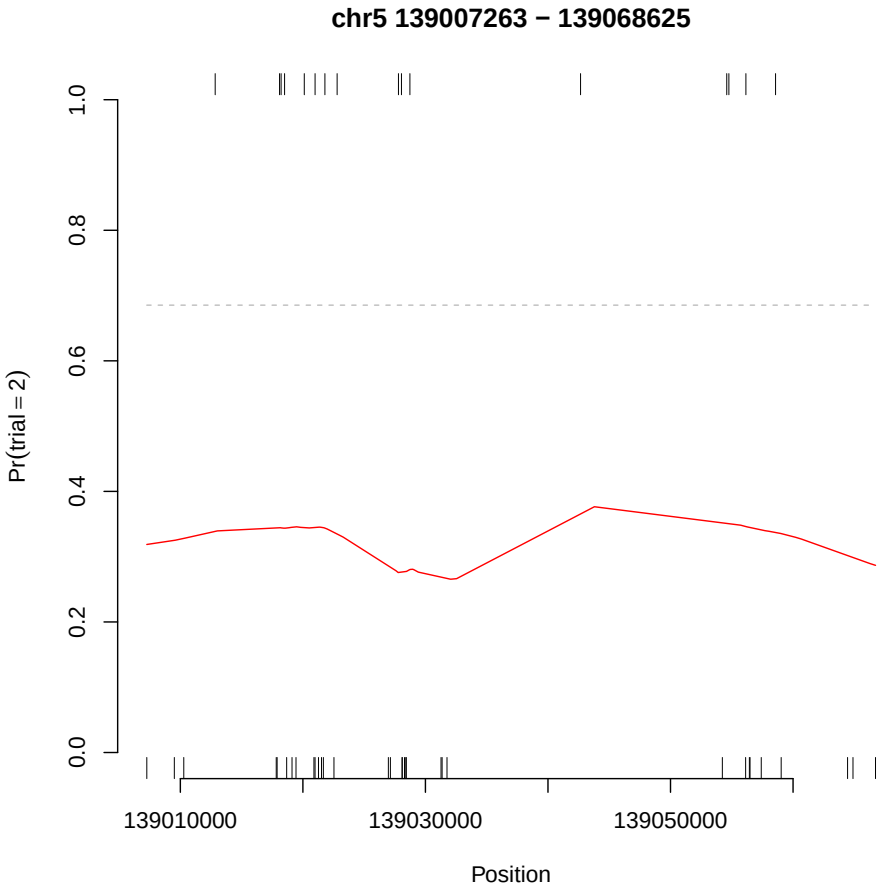
5.15 Clump chr6:15375728-15472933

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr6	15375728	97206	17	2	61	-2.72	0.037



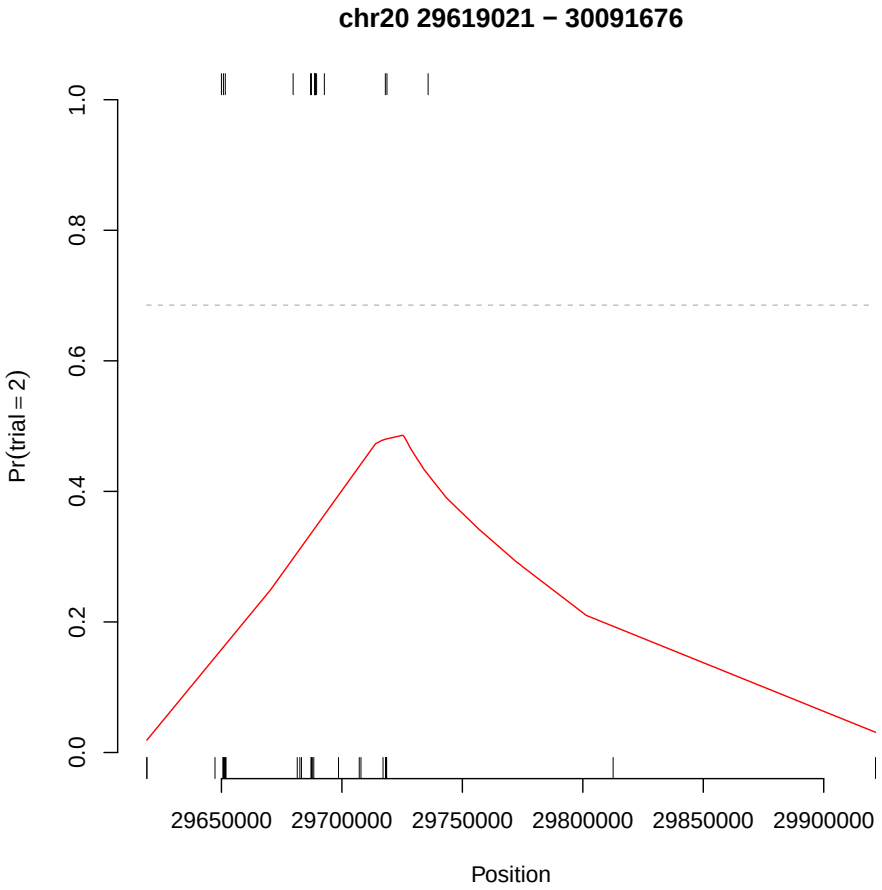
5.16 Clump chr5:139007263-139068625

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr5	139007263	61363	33	16	54	-1.49	0.037



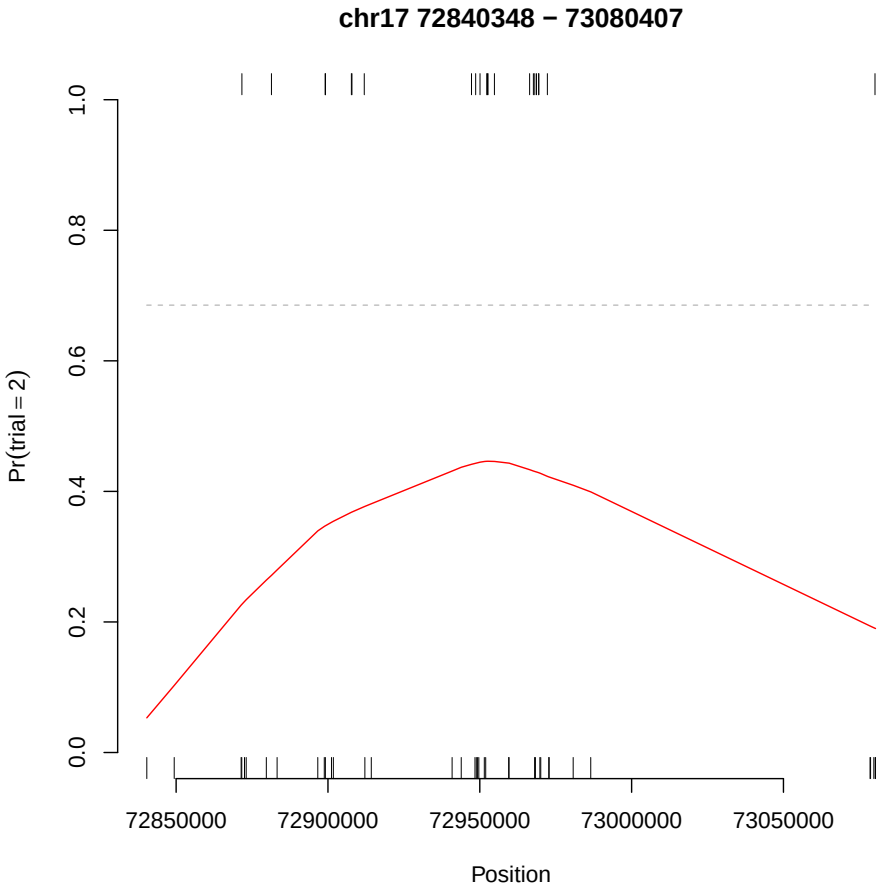
5.17 Clump chr20:29619021-30091676

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr20	29619021	472656	32	15	56	-1.52	0.037



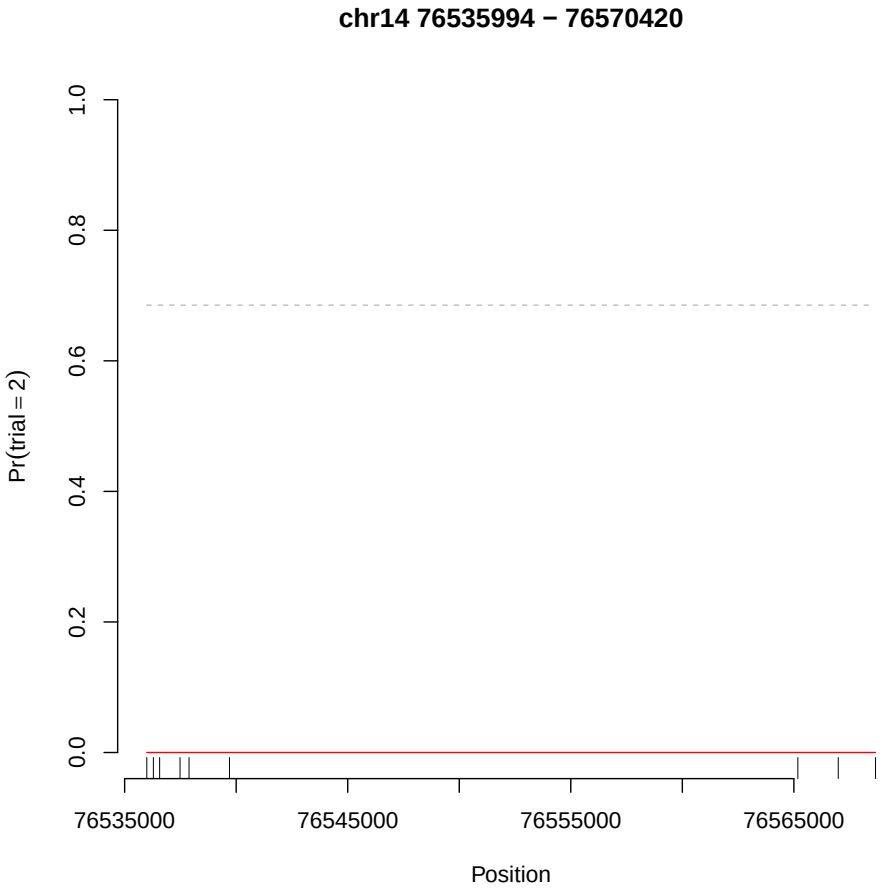
5.18 Clump chr17:72840348-73080407

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr17	72840348	240060	38	22	49	-1.31	0.0393



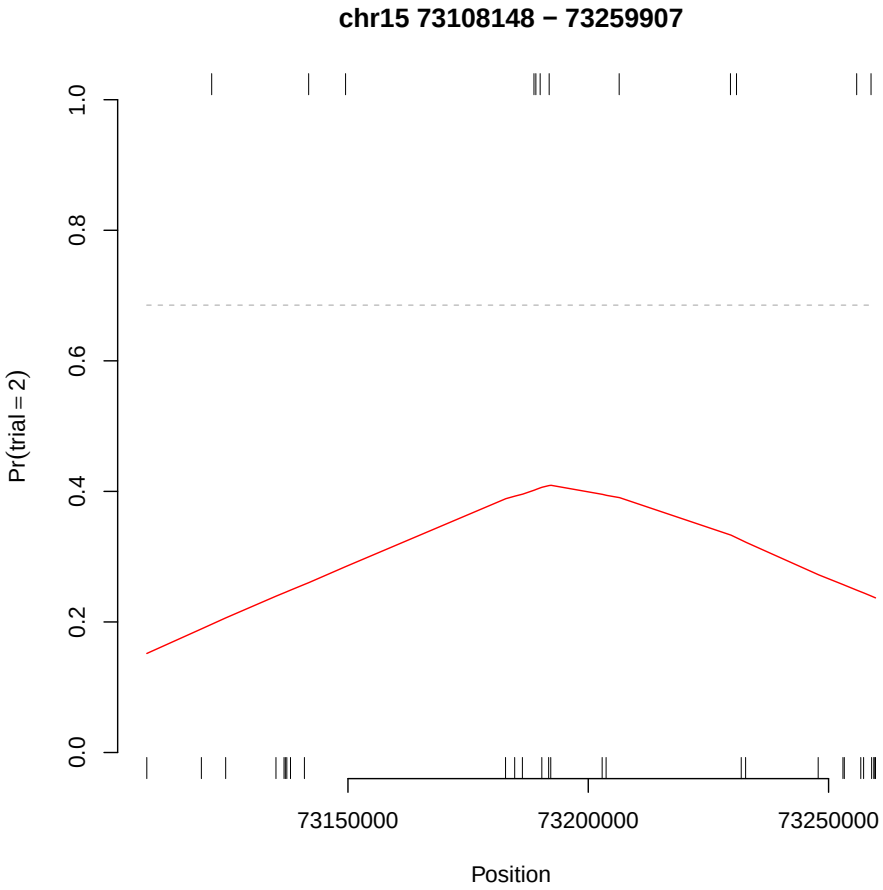
5.19 Clump chr14:76535994-76570420

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr14	76535994	34427	9	0	59	-3.72	0.0395



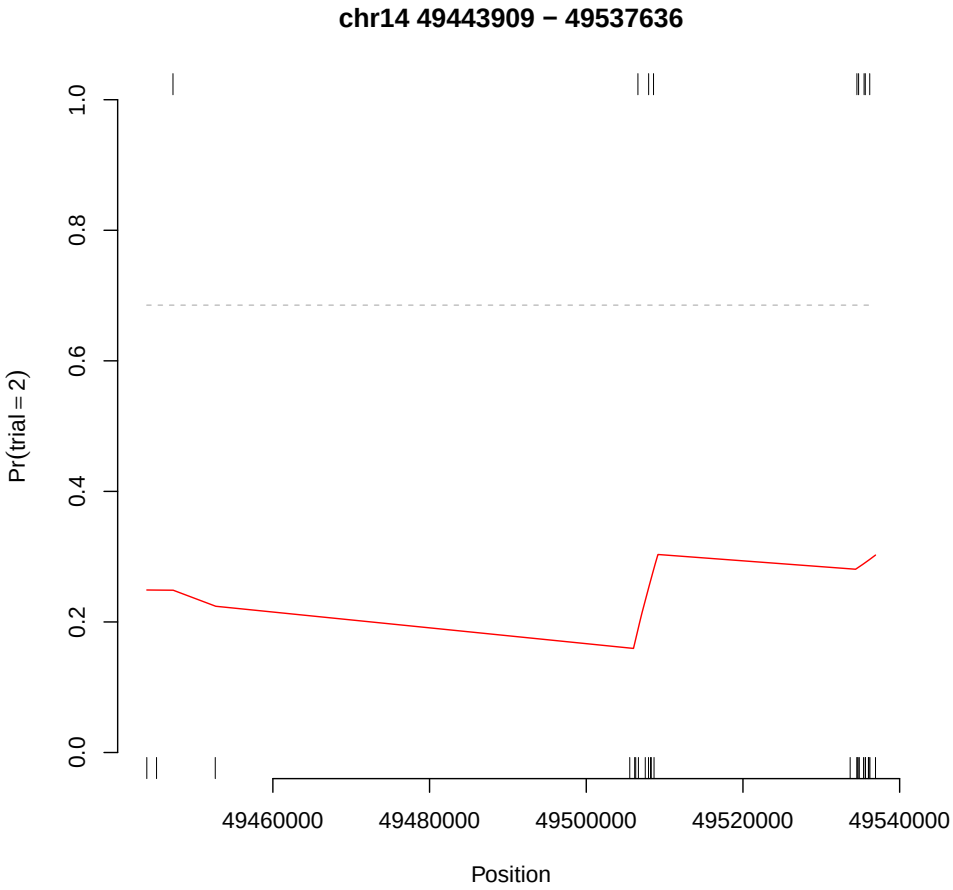
5.20 Clump chr15:73108148-73259907

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr15	73108148	151760	28	12	51	-1.6	0.0395



5.21 Clump chr14:49443909-49537636

Chromosome	start	width	SCID1	SCID2	depth	log.OR	FDR
chr14	49443909	93728	25	9	57	-1.77	0.0395



6 Software

This report was prepared using the following software:

R [R Core Team, 2012]

BioConductor:: [Gentleman et al., 2004] packages:

- GenomicRanges [Aboyoun et al., 2012]
- BSgenome [Pages, 2012]
- and its supporting infrastructure.

R packages

- knitr [Xie, 2012]
- brew [Horner, 2011]
- geneRxCluster [Berry, 2014] available at `http://www.bioconductor.org/`

Emacs

- orgmode [Dominik, 2010] was used to prepare the knitr and brew scripts and to manage data preparation.

References

[Aboyoun et al., 2012] Aboyoun, P., Pages, H., and Lawrence, M. (2012). *GenomicRanges: Representation and manipulation of genomic intervals*. R package version 1.8.3.

[Aldous, 2010] Aldous, D. (2010). *Probability approximations via the Poisson clumping heuristic*. Springer-Verlag New York, Inc.

[Berry, 2014] Berry, C. (2014). *geneRxCluster: gRx Differential Clustering*. R package version 1.0.0.

[Berry et al., 2014] Berry, C. C., Ocwieja, K. E., Malani, N., and Bushman, F. D. (2014). Comparing DNA site clusters with Scan Statistics. *Bioinformatics*.

[Dominik, 2010] Dominik, C. (2010). *The Org-Mode 7 Reference Manual: Organize Your Life with GNU Emacs*. Network Theory, UK. with contributions by David O'Toole, Bastien Guerry, Philip Rooke, Dan Davison, Eric Schulte, and Thomas Dye.

[Gentleman et al., 2004] Gentleman, R. C., Carey, V. J., Bates, D. M., and others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

[Horner, 2011] Horner, J. (2011). *brew: Templating Framework for Report Generation*. R package version 1.0-6.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Pages, 2012] Pages, H. (2012). *BSgenome: Infrastructure for Biostrings-based genome data packages*. R package version 1.24.0.

[R Core Team, 2012] R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Siegmund et al., 2011] Siegmund, D., Zhang, N., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

[Xie, 2012] Xie, Y. (2012). *knitr: A general-purpose package for dynamic report generation in R*. R package version 0.7.