Multimodal Data Integration for Cancer Virus Discovery

Matthew Wyczalkowski
Ding Lab

TCGA Symposium
May 13, 2014

Introduction

- 10-15% of cancers worldwide caused by viruses.
- Several virus types are associated with human cancers.
    - Role of other viruses remains controversial.
- Some viruses integrate into the host genome, others remain episomal.
    - What is role of integration in gene expression?
    - How is integration associated with disease initiation and progression?
- Viruses are ubiquitous, cancer is not. Why?

TCGA Samples by Disease and Type

Unique barcodes in CGHUB

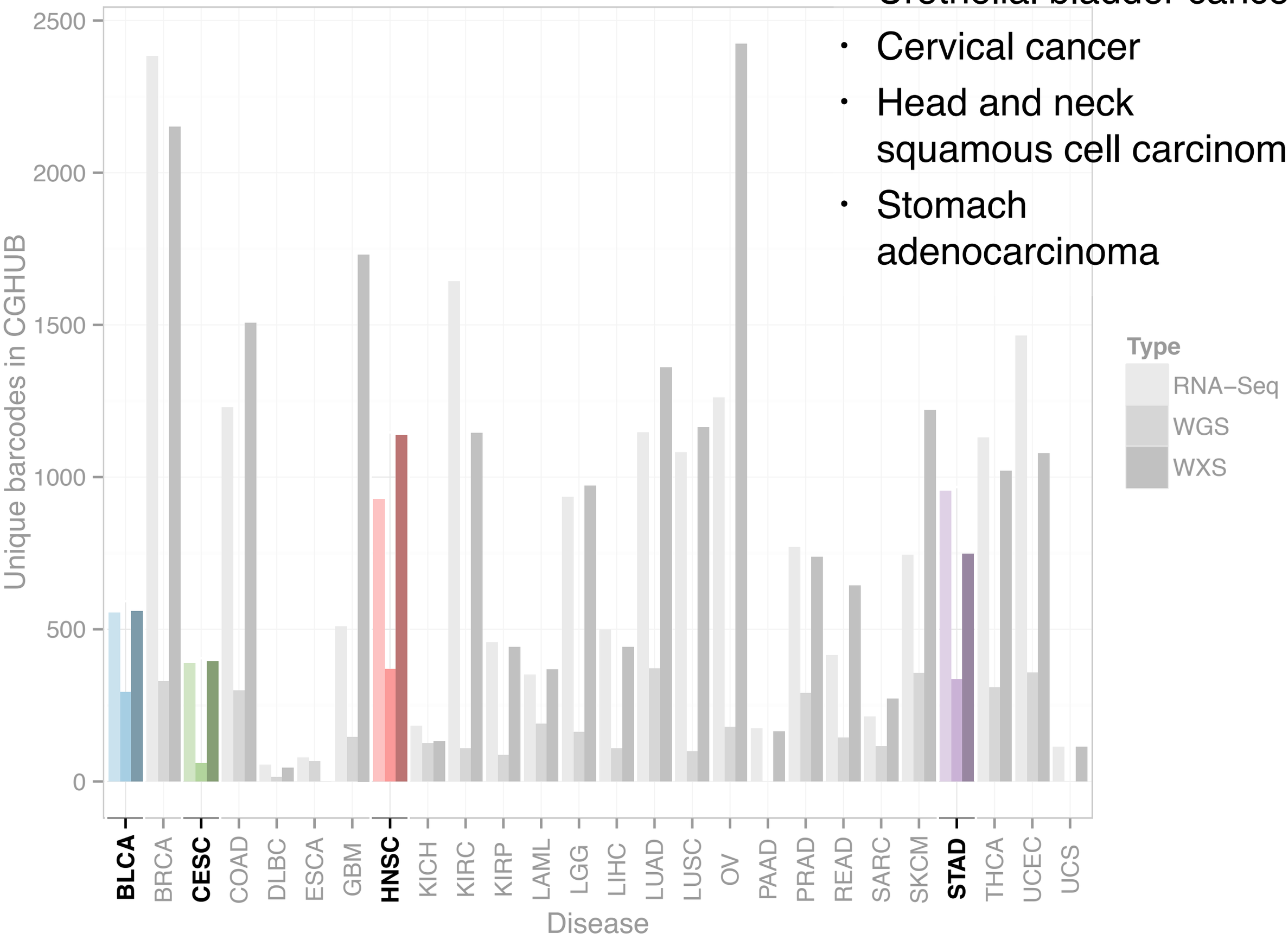
2500
2000
1500
1000
500
0

BLCA BRCA CESC COAD DLBC ESCA GBM HNSC KICH KIRC KIRP LAML LGG LIHC LUAD LUSC OV PAAD PRAD READ SARC SKCM STAD THCA UCEC UCS

Disease

Type
RNA–Seq
WGS
WXS

- Urothelial bladder cancer
- Cervical cancer
- Head and neck squamous cell carcinoma
- Stomach adenocarcinoma



Virus Analysis Pipeline

RNA-Seq
(Human reference)

Extract Unmapped Reads

Blast to NT Database

Virus Discovery

Select Virus Reference

Virus Analysis Pipeline

RNA-Seq
(Human reference)

Extract Unmapped Reads

Blast to NT Database

Virus Discovery

Select Virus Reference

WGS
(Human reference)

WGS
(Human+ select virus reference)

Virus Analysis Pipeline

RNA-Seq
(Human reference)

Extract Unmapped Reads

Blast to NT Database

Virus Discovery

Select Virus Reference

WGS
(Human reference)

WGS
(Human+ select virus reference)

CESC Exome Sequencing
(with HPV probes)

WXS
(Human+HPV reference)

Virus Analysis Pipeline

RNA-Seq
(Human reference)

Extract Unmapped Reads

Blast to NT Database

Virus Discovery

Select Virus Reference

WGS
(Human reference)

WGS
(Human+ select virus reference)

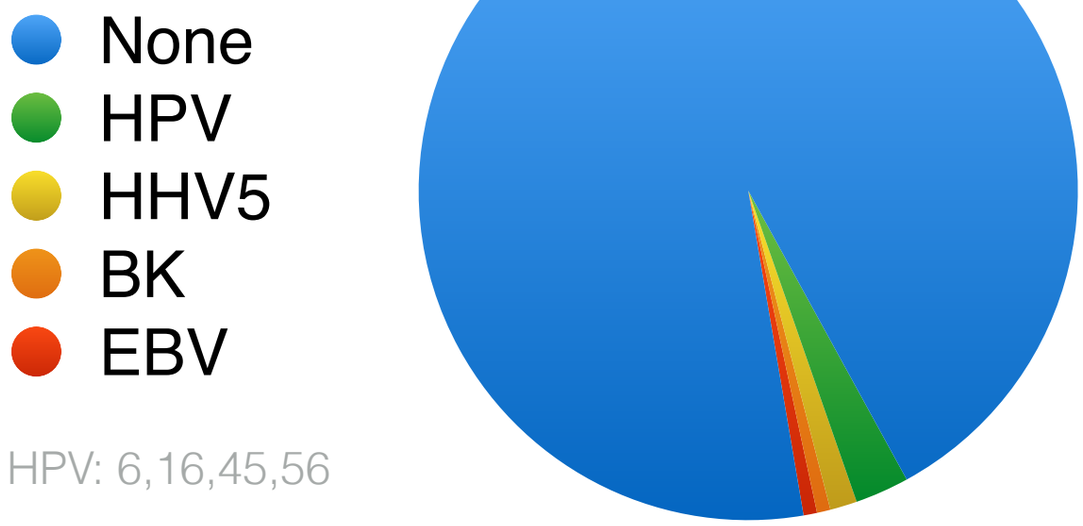
CESC Exome Sequencing
(with HPV probes)

WXS
(Human+HPV reference)

Viral Status and Integration

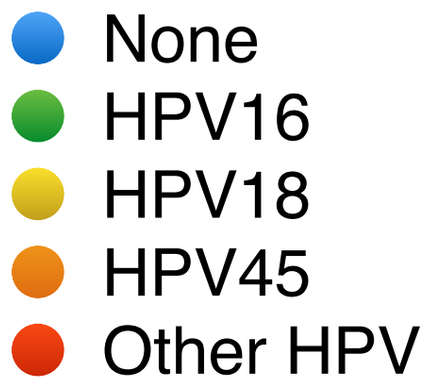
Viral Patterns

BLCA

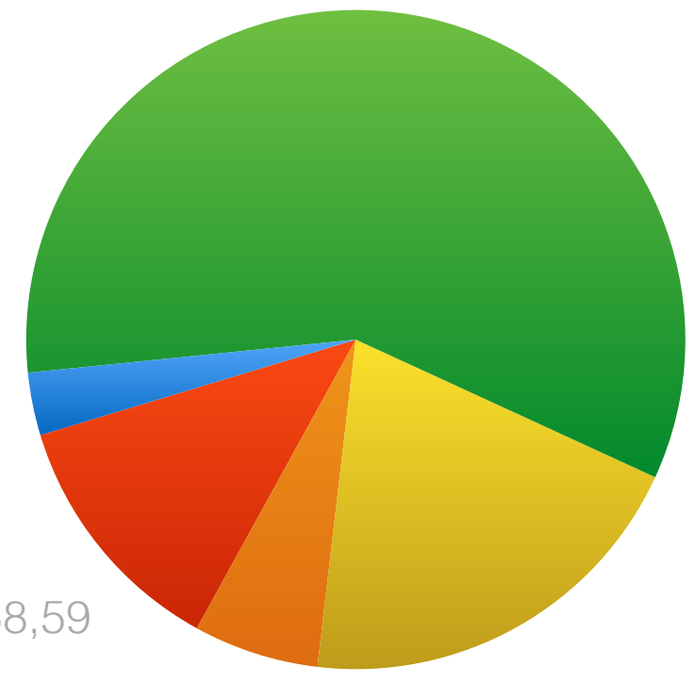


HPV: 6,16,45,56

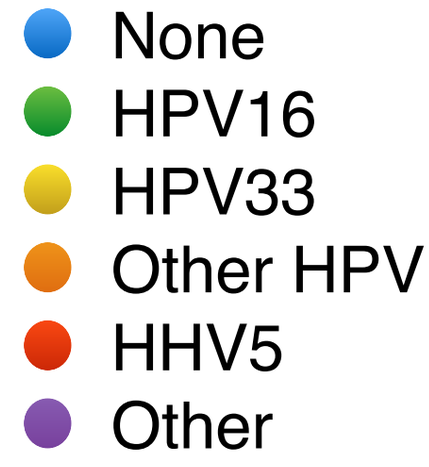
CESC



Other HPV: 31,33,39,52,58,59



HNSC

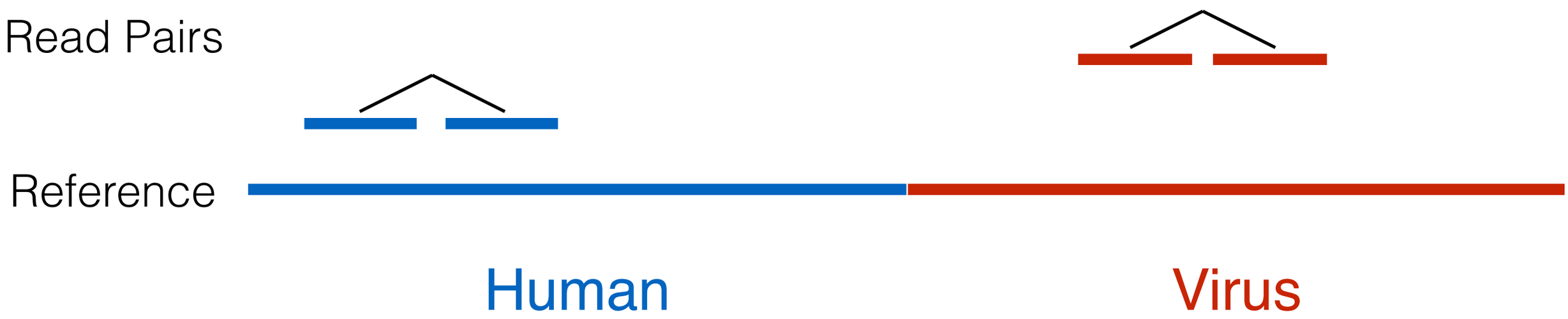


Other HPV: 35,56
Other: EBV, HSV1

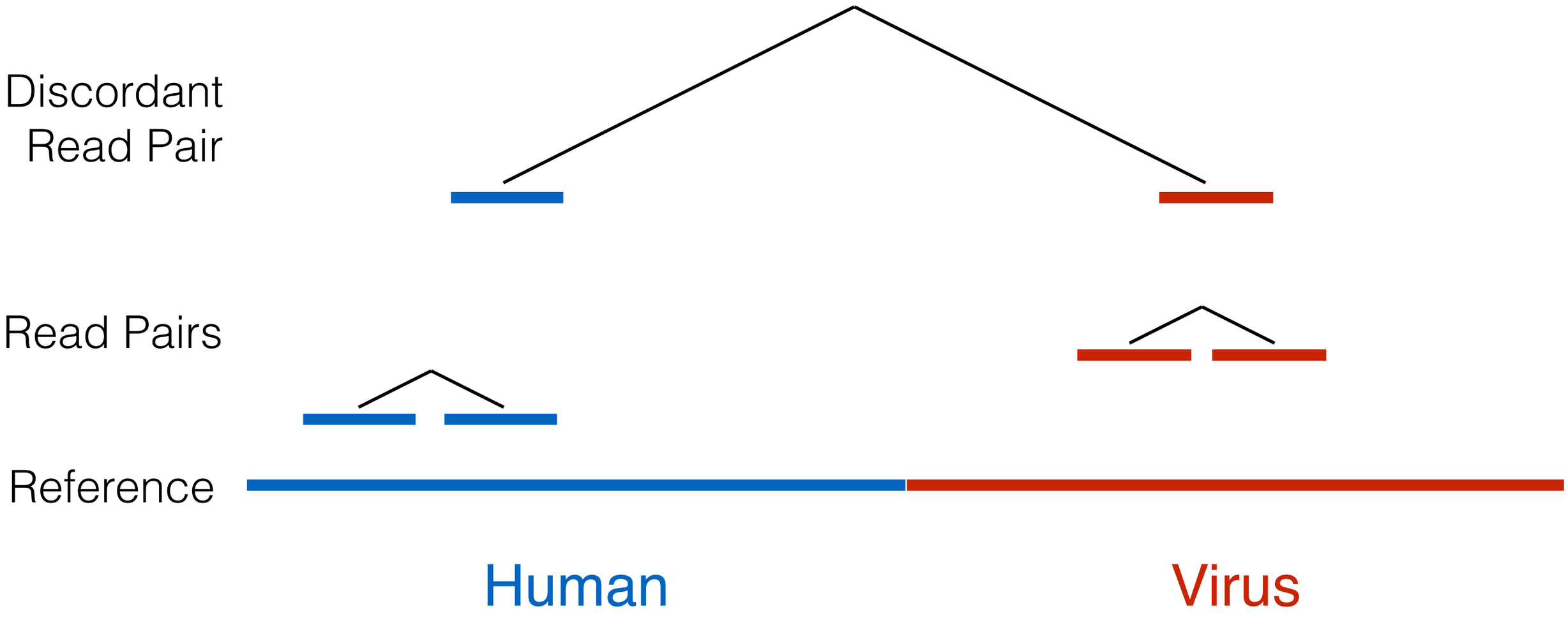
STAD

Integration

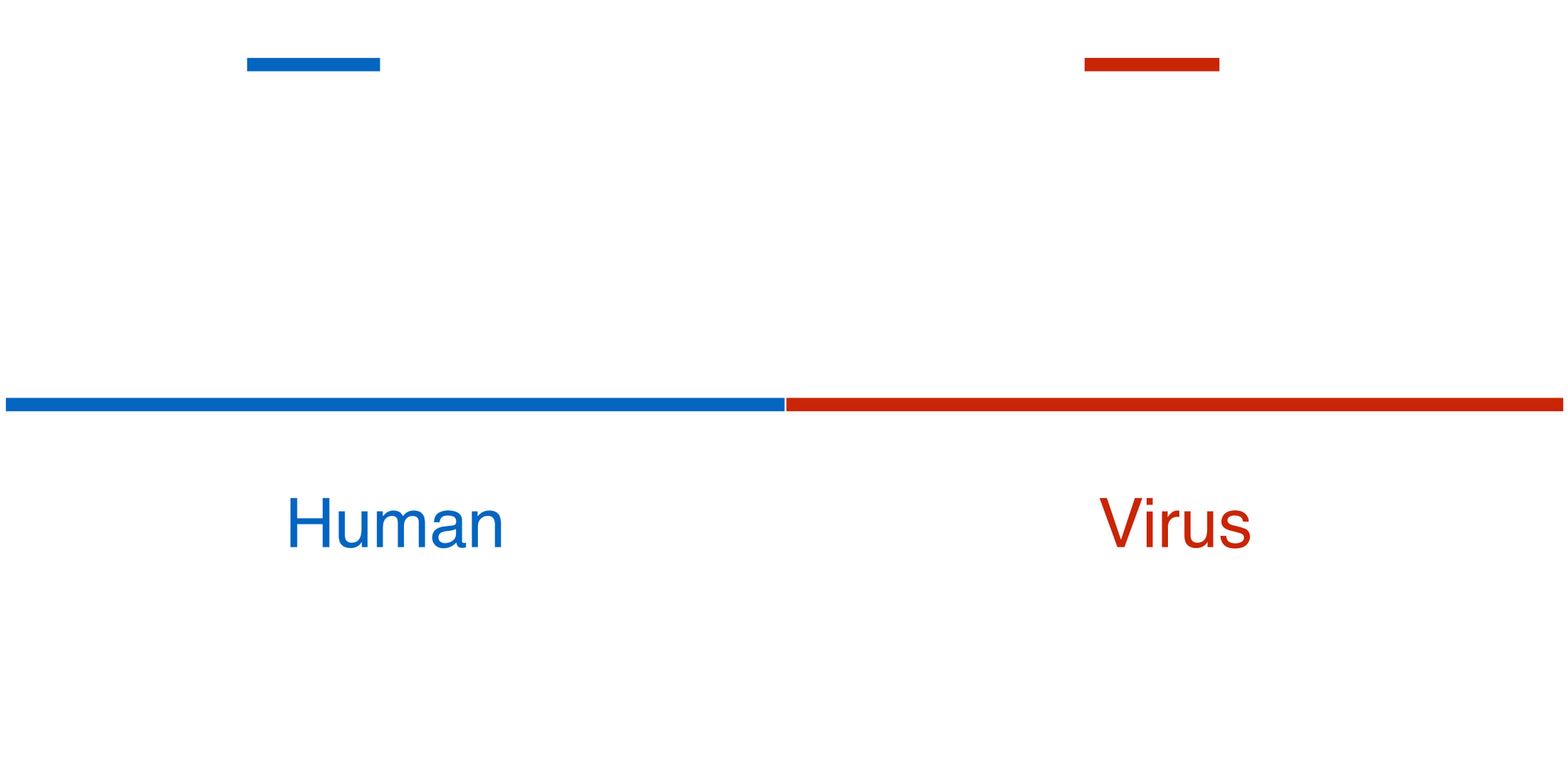
Integration



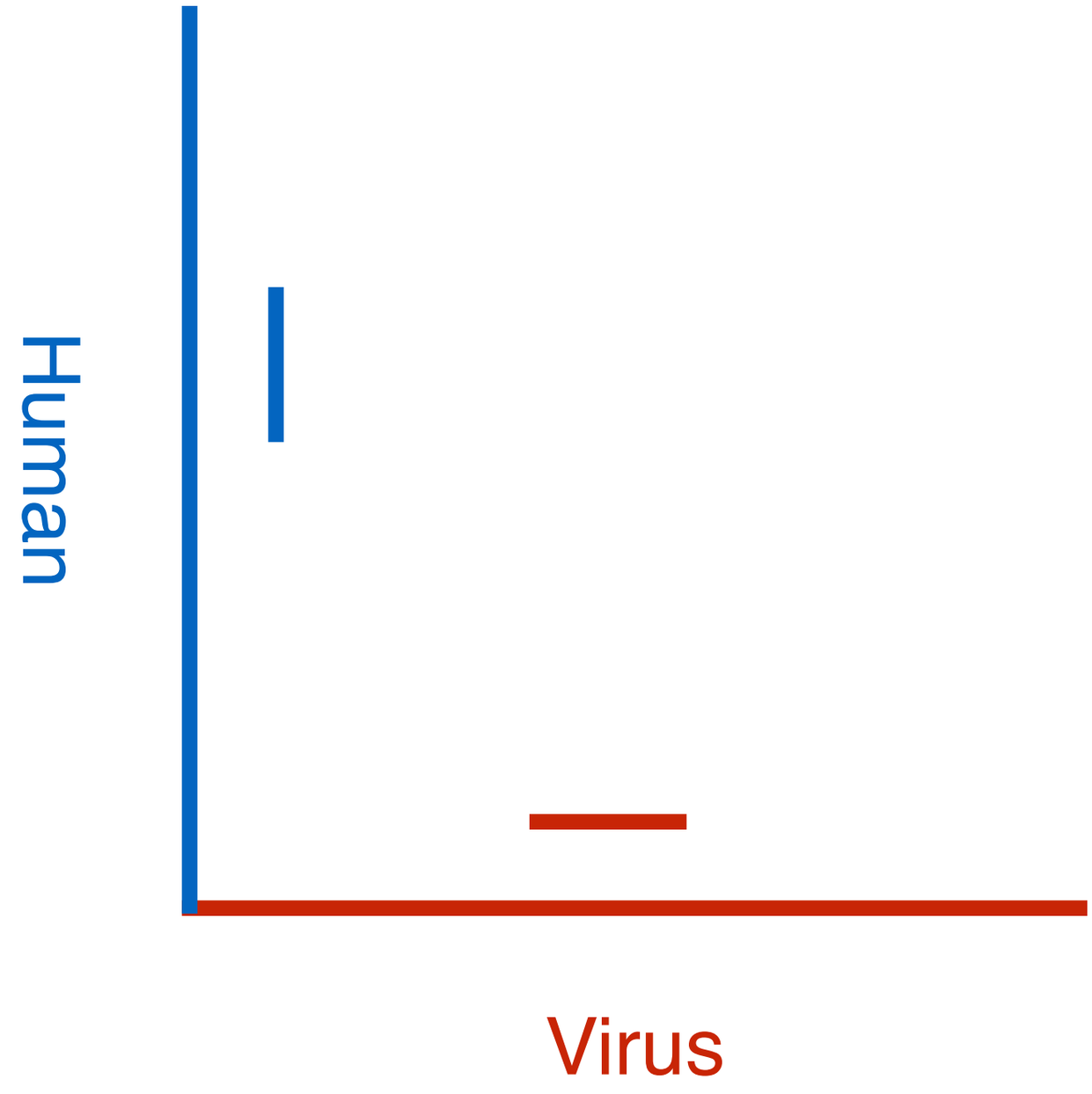
Integration



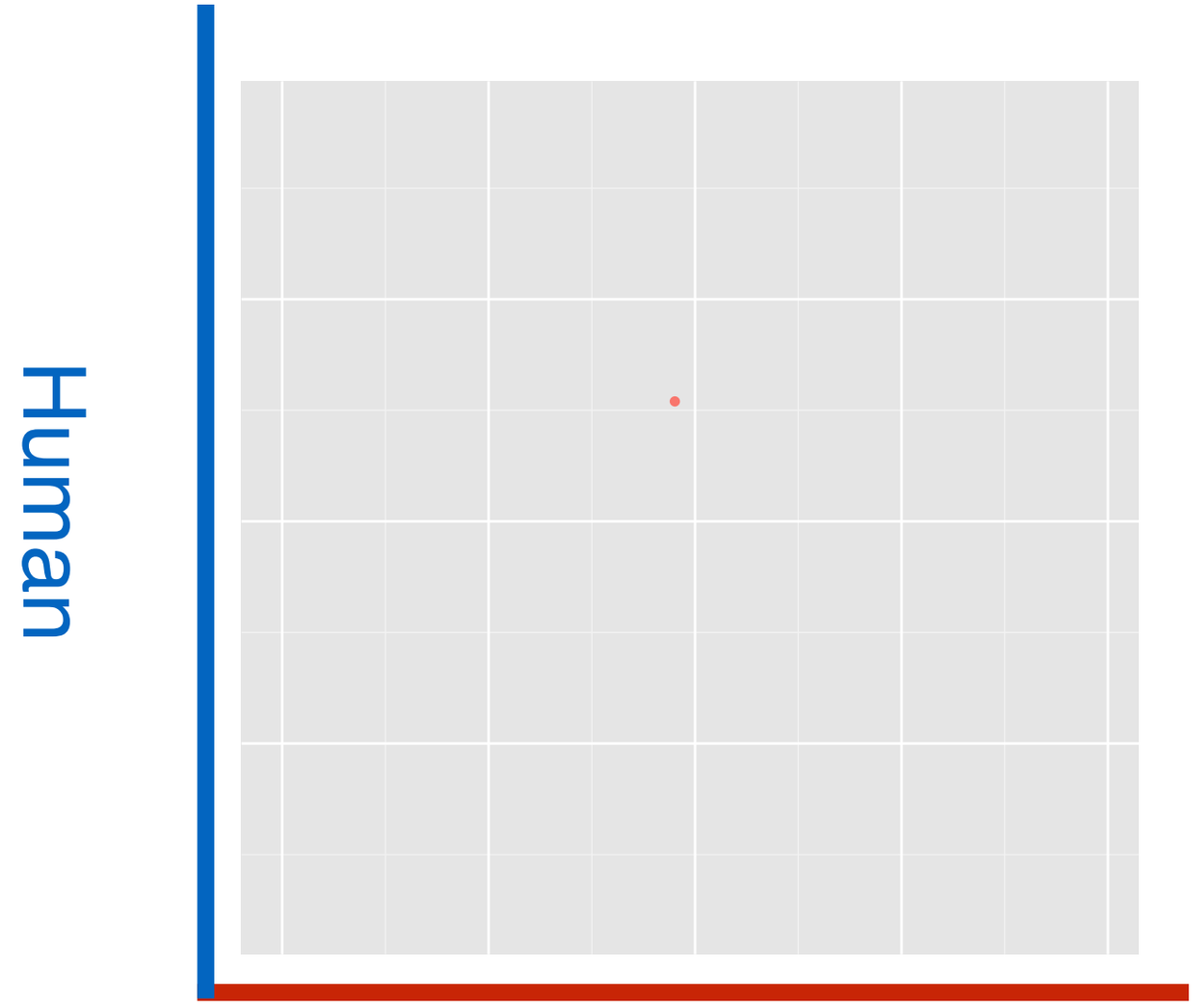
Integration



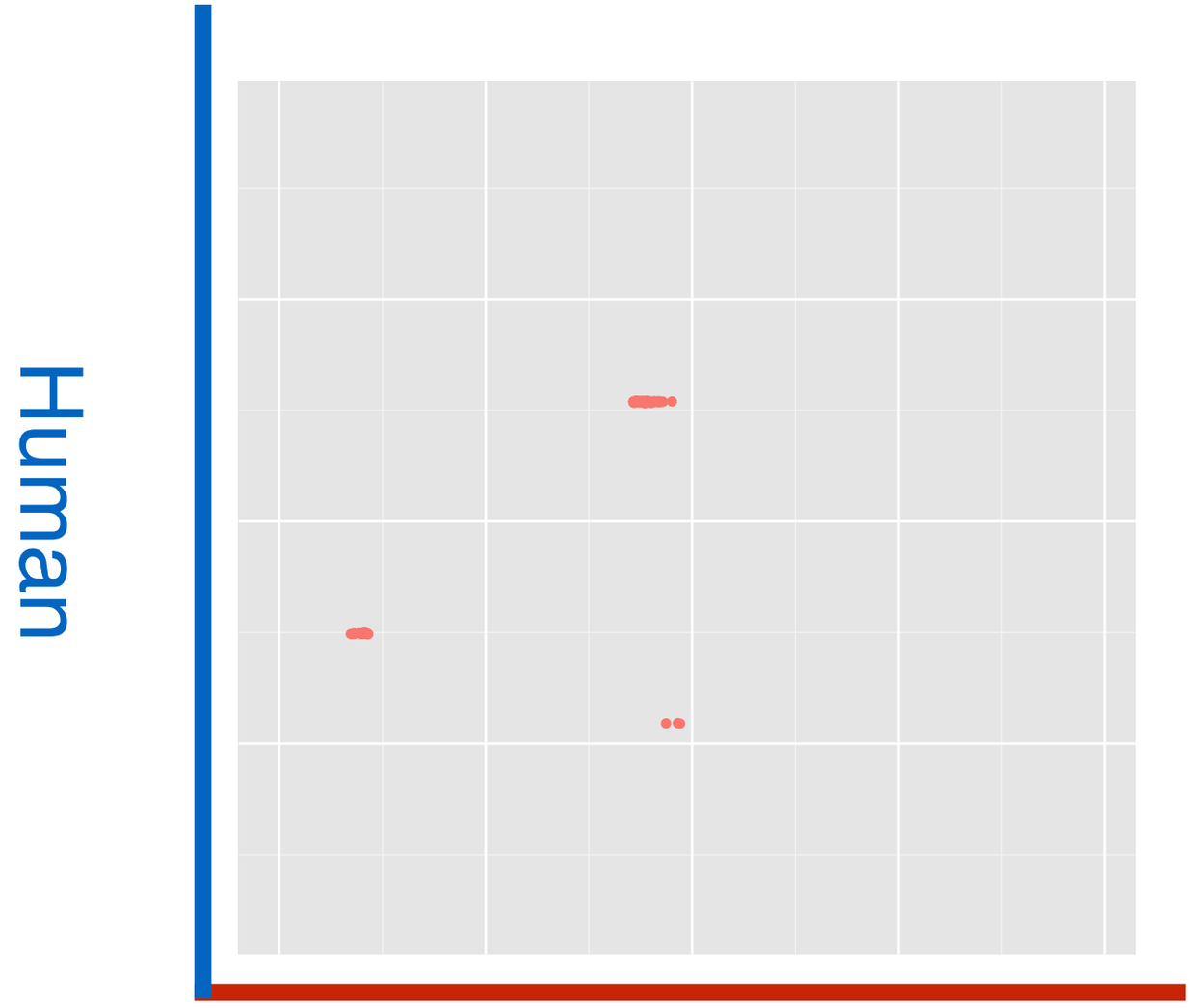
Integration



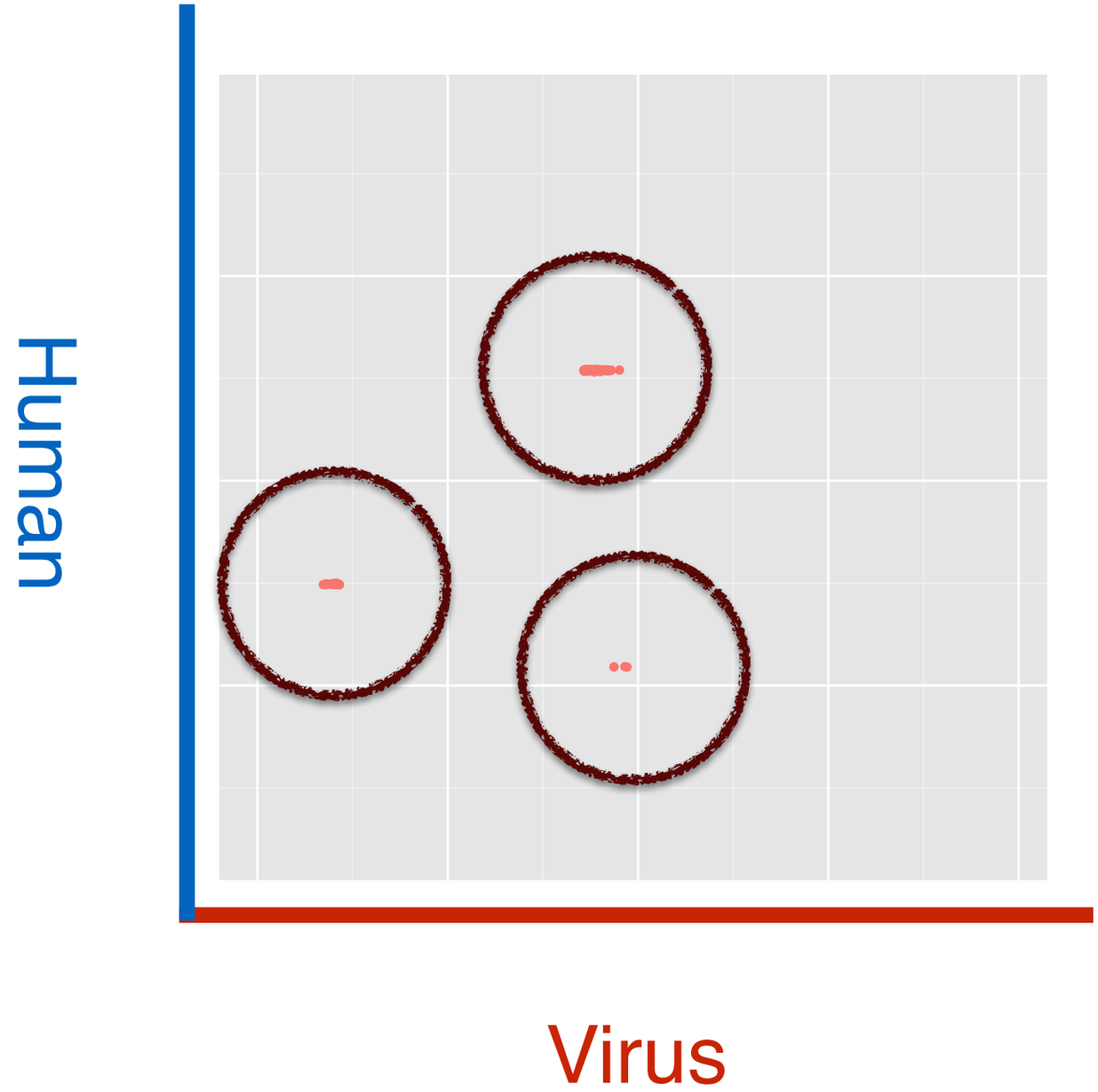
Integration



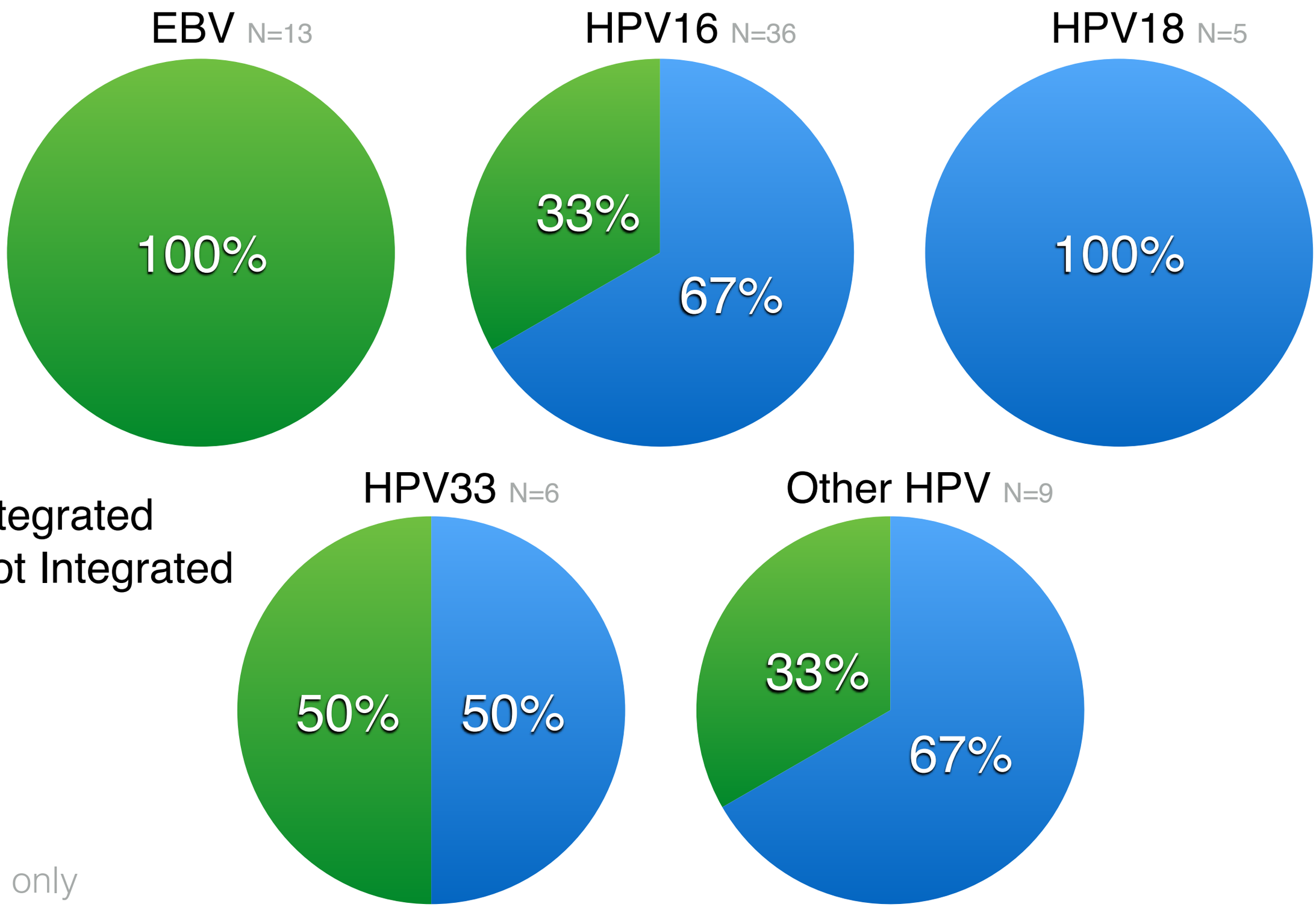
Integration



Integration



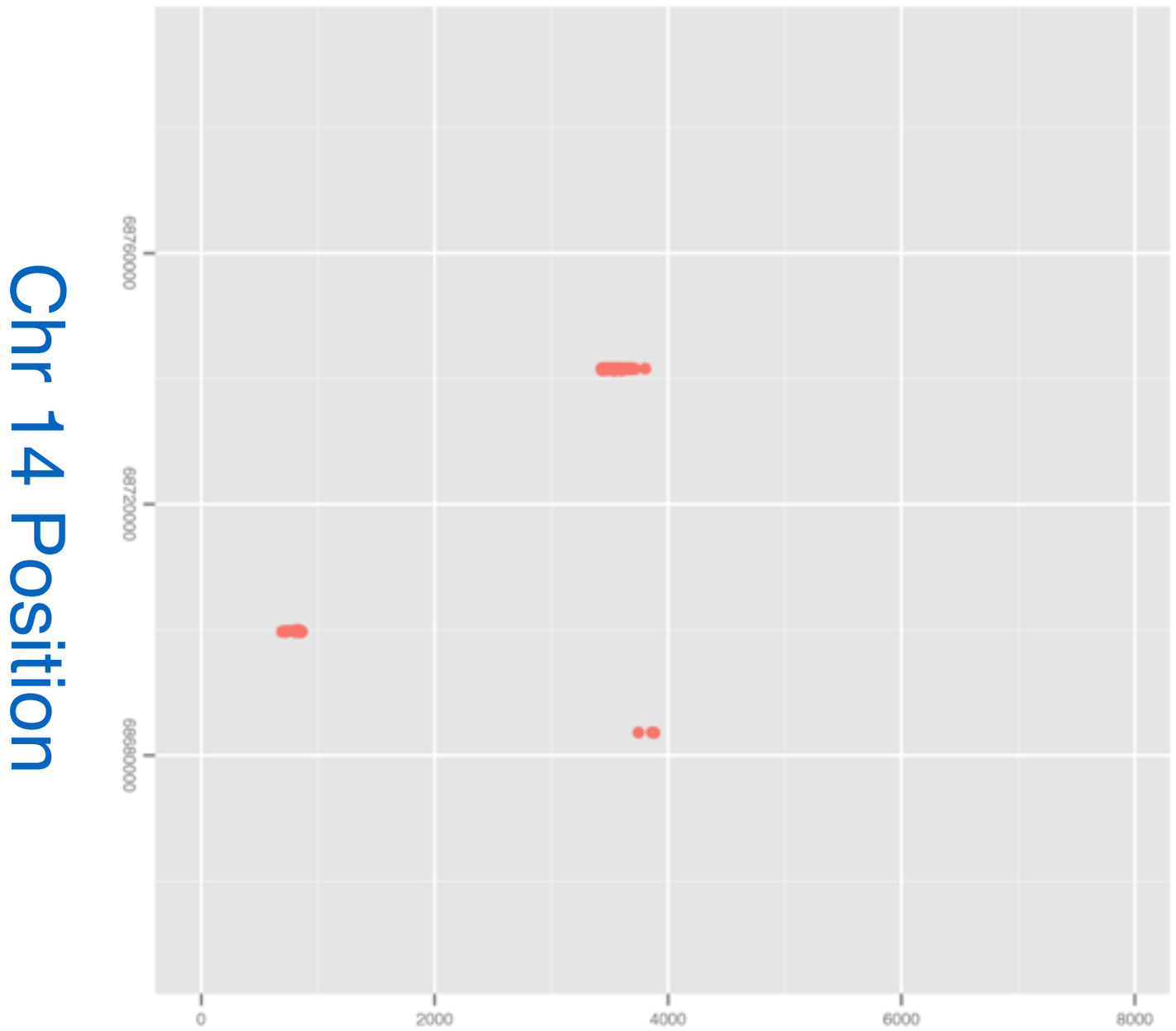
Integration Patterns



Head and Neck

BA-4077-01

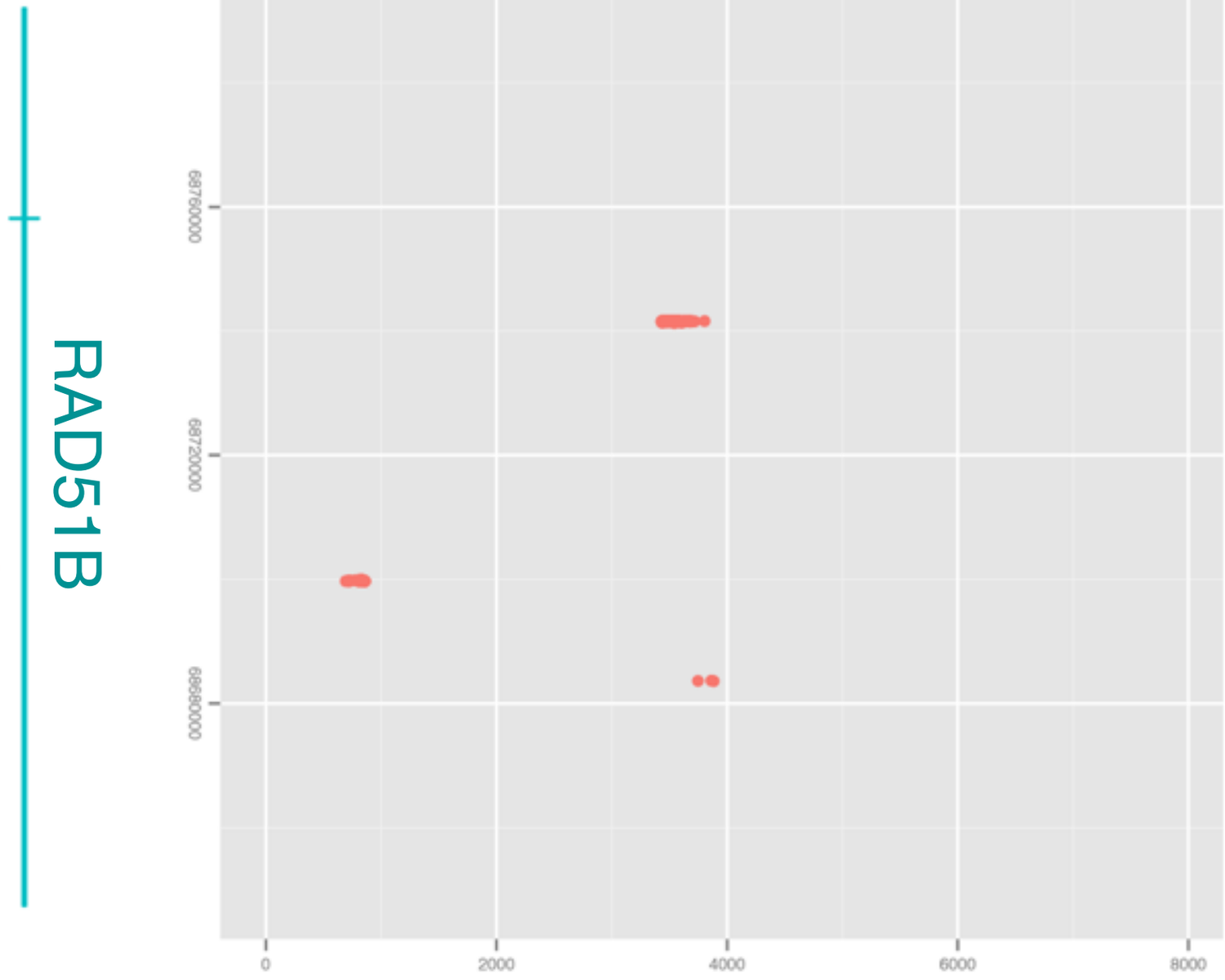
118 Discordant Pairs



Head and Neck

BA-4077-01

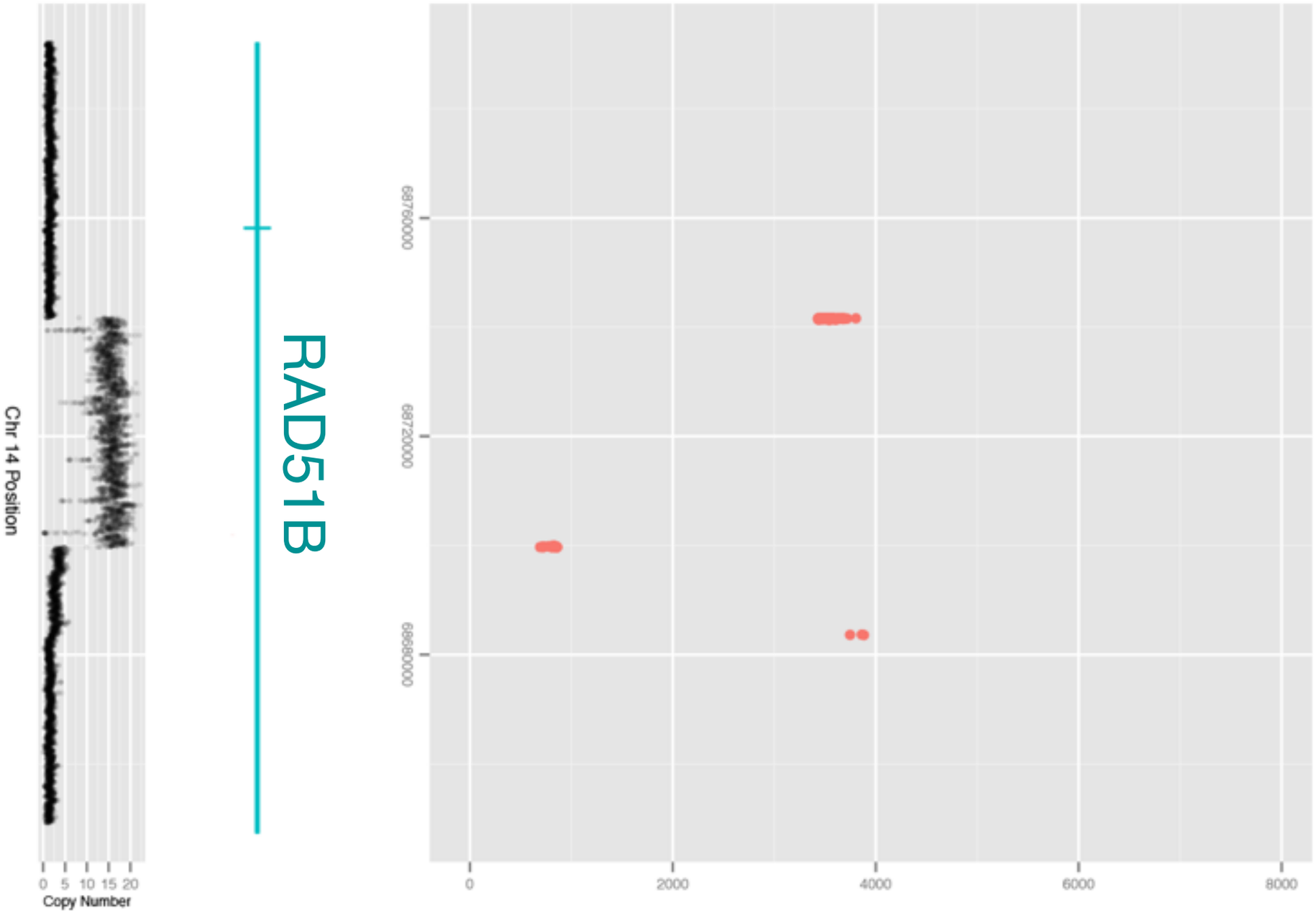
118 Discordant Pairs



Head and Neck

BA-4077-01

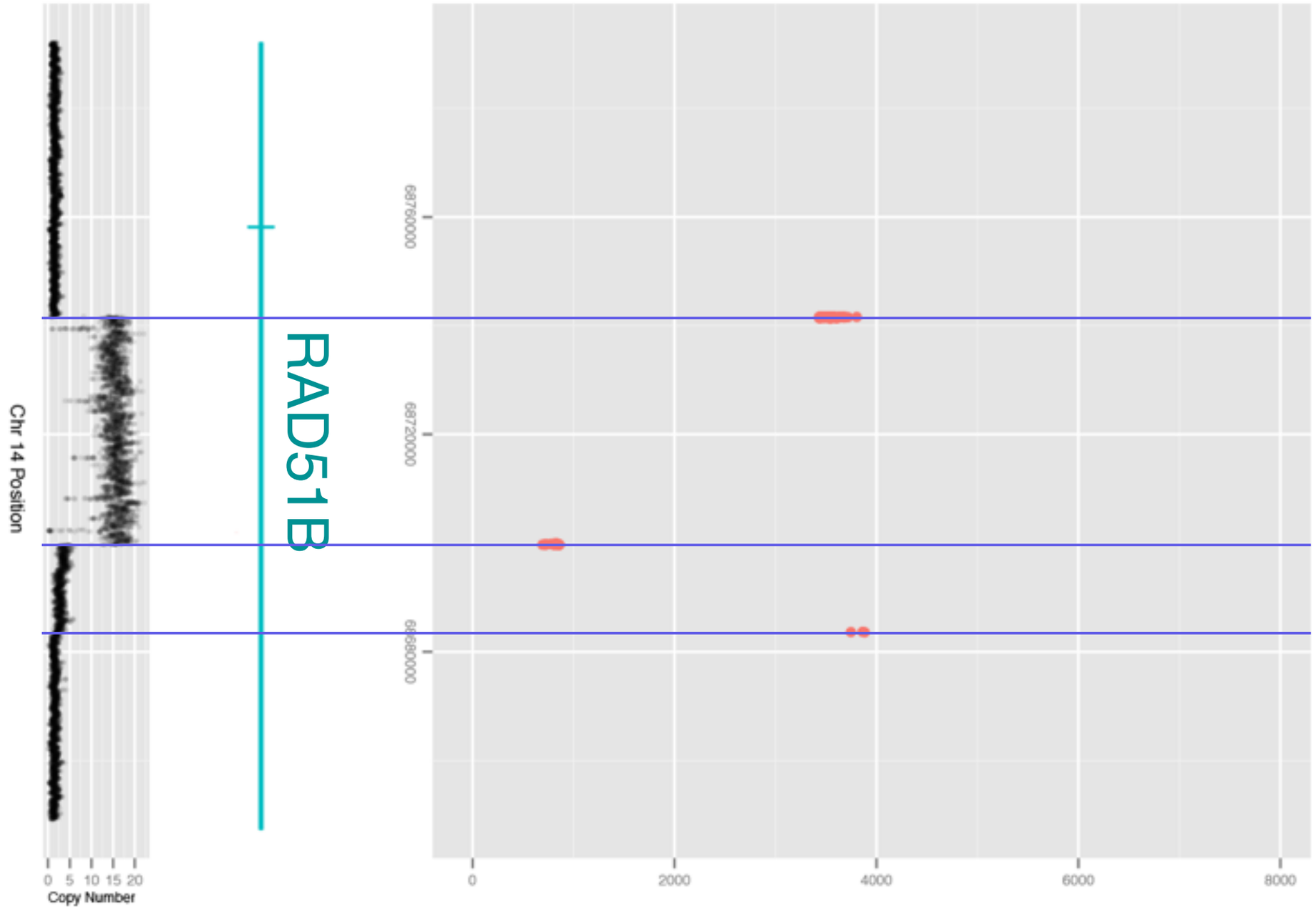
118 Discordant Pairs



Head and Neck

BA-4077-01

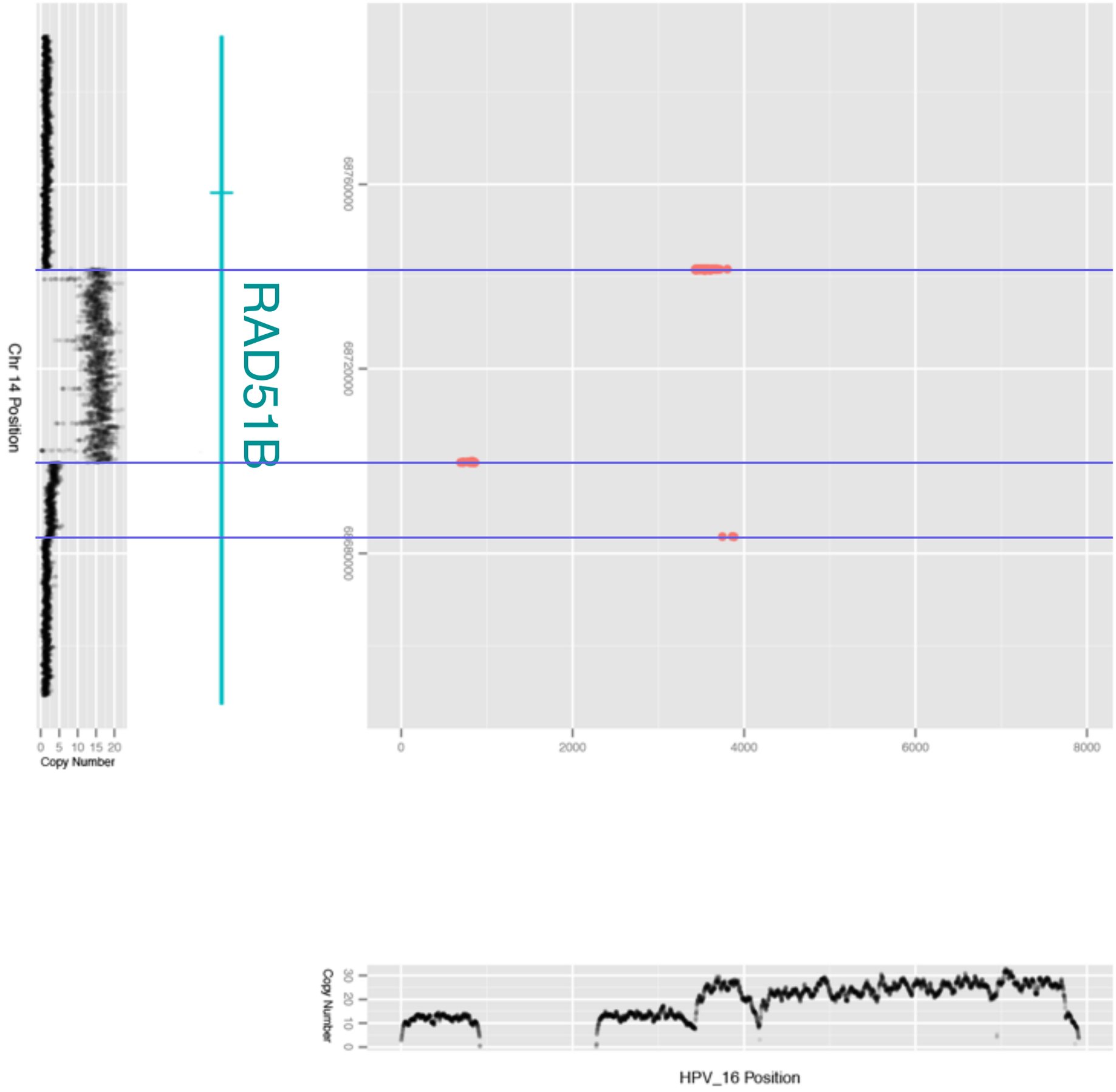
118 Discordant Pairs



Head and Neck

BA-4077-01

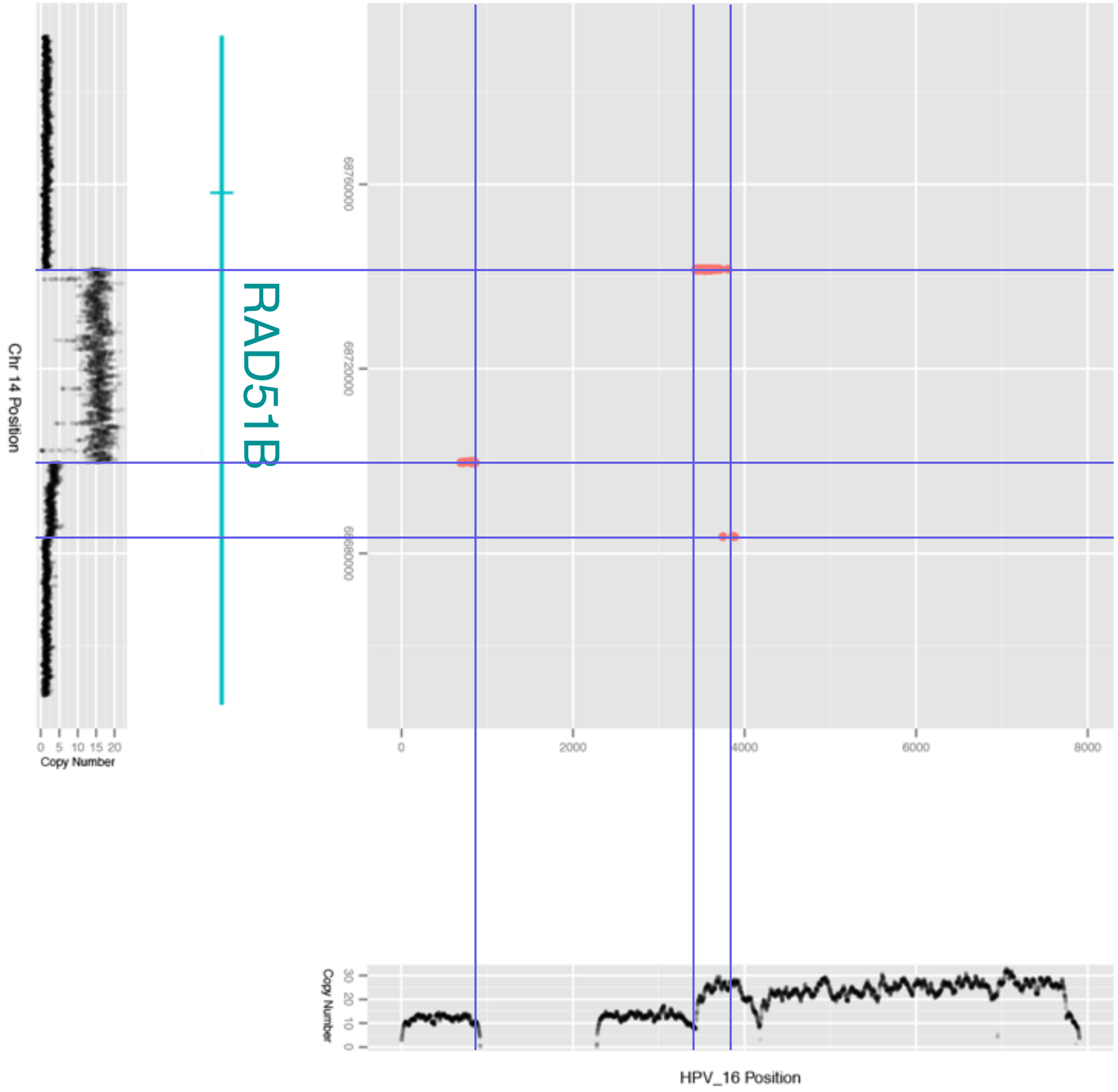
118 Discordant Pairs



Head and Neck

BA-4077-01

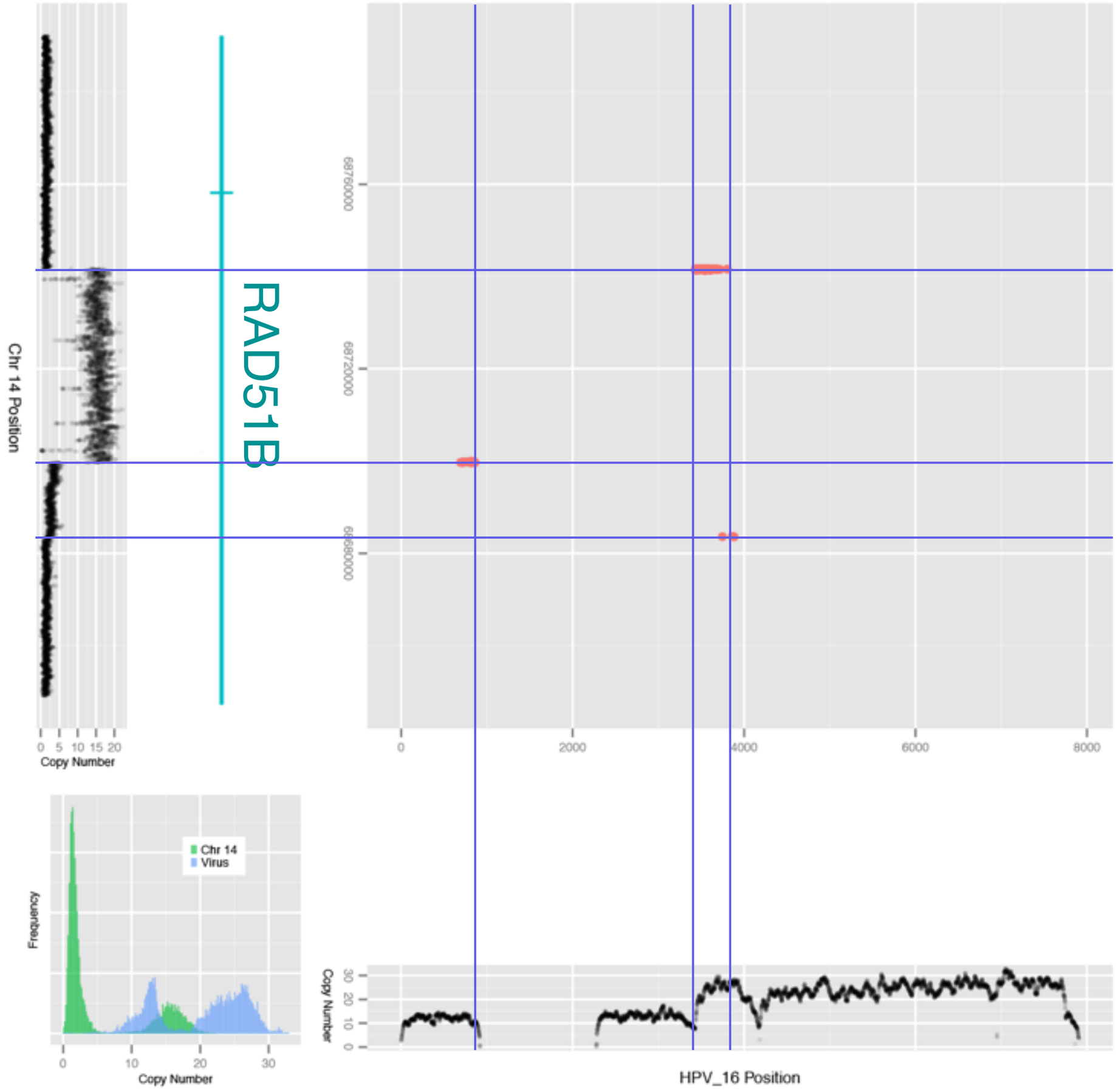
118 Discordant Pairs



Head and Neck

BA-4077-01

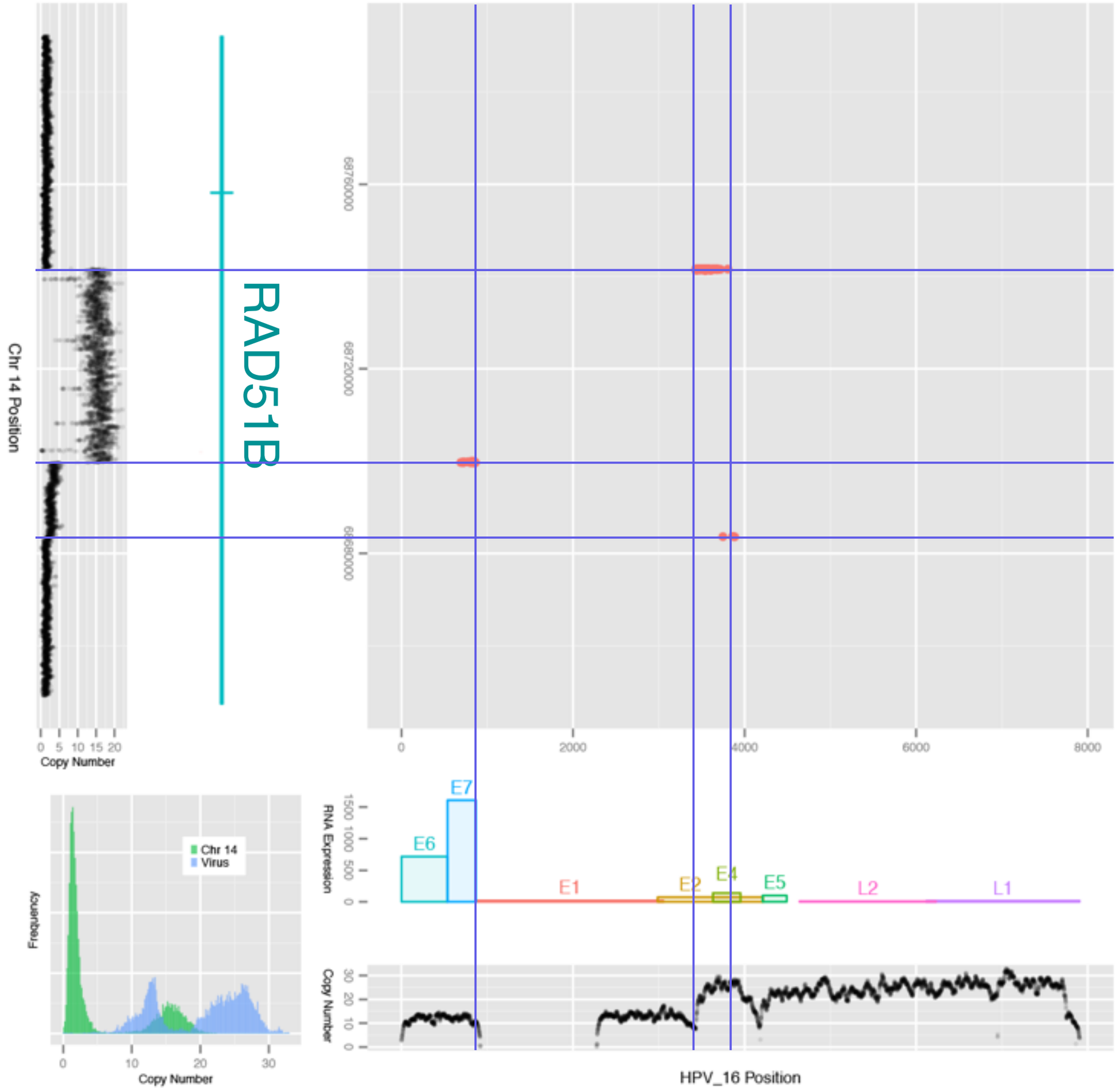
118 Discordant Pairs



Head and Neck

BA-4077-01

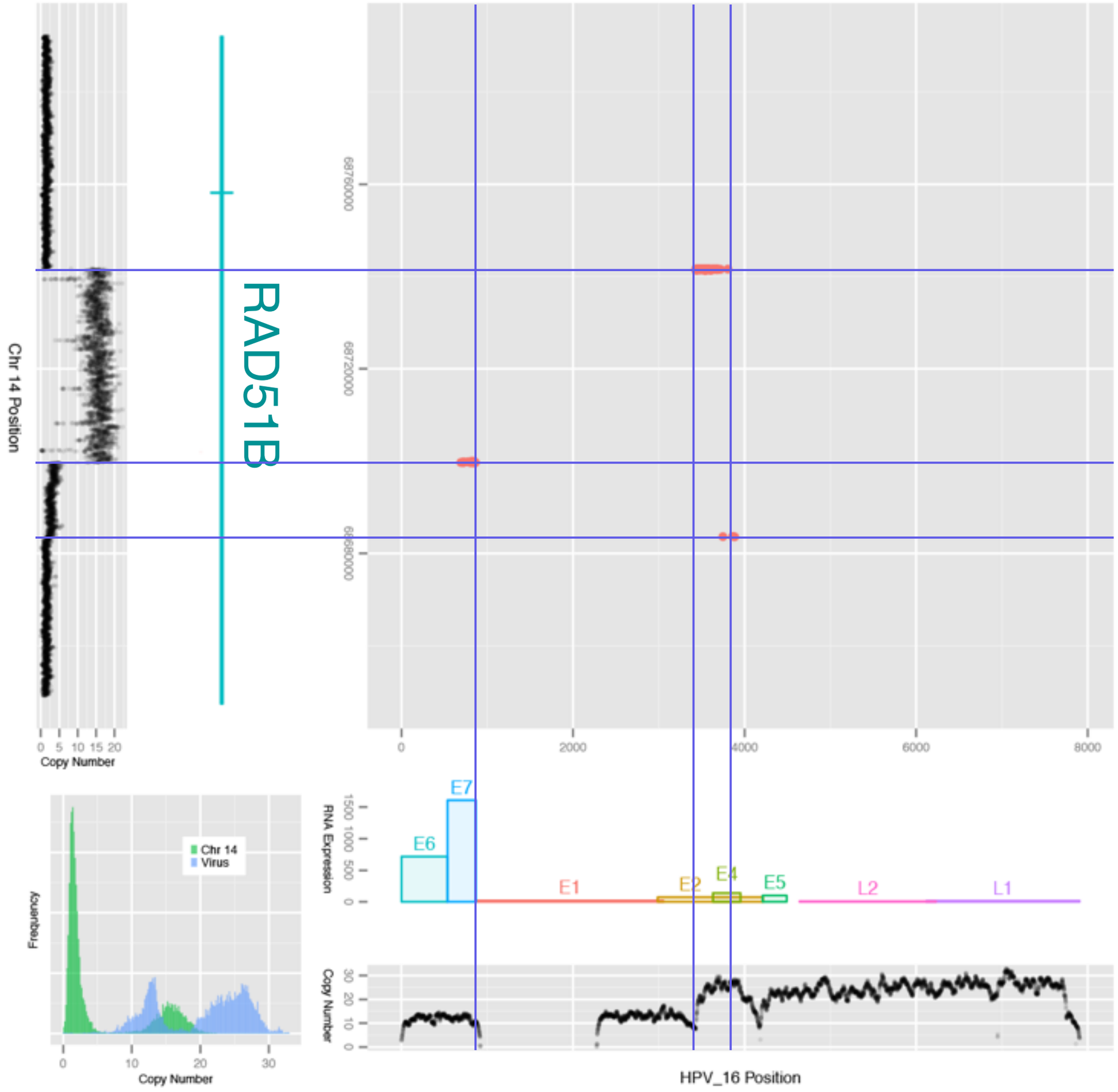
118 Discordant Pairs



Head and Neck

BA-4077-01

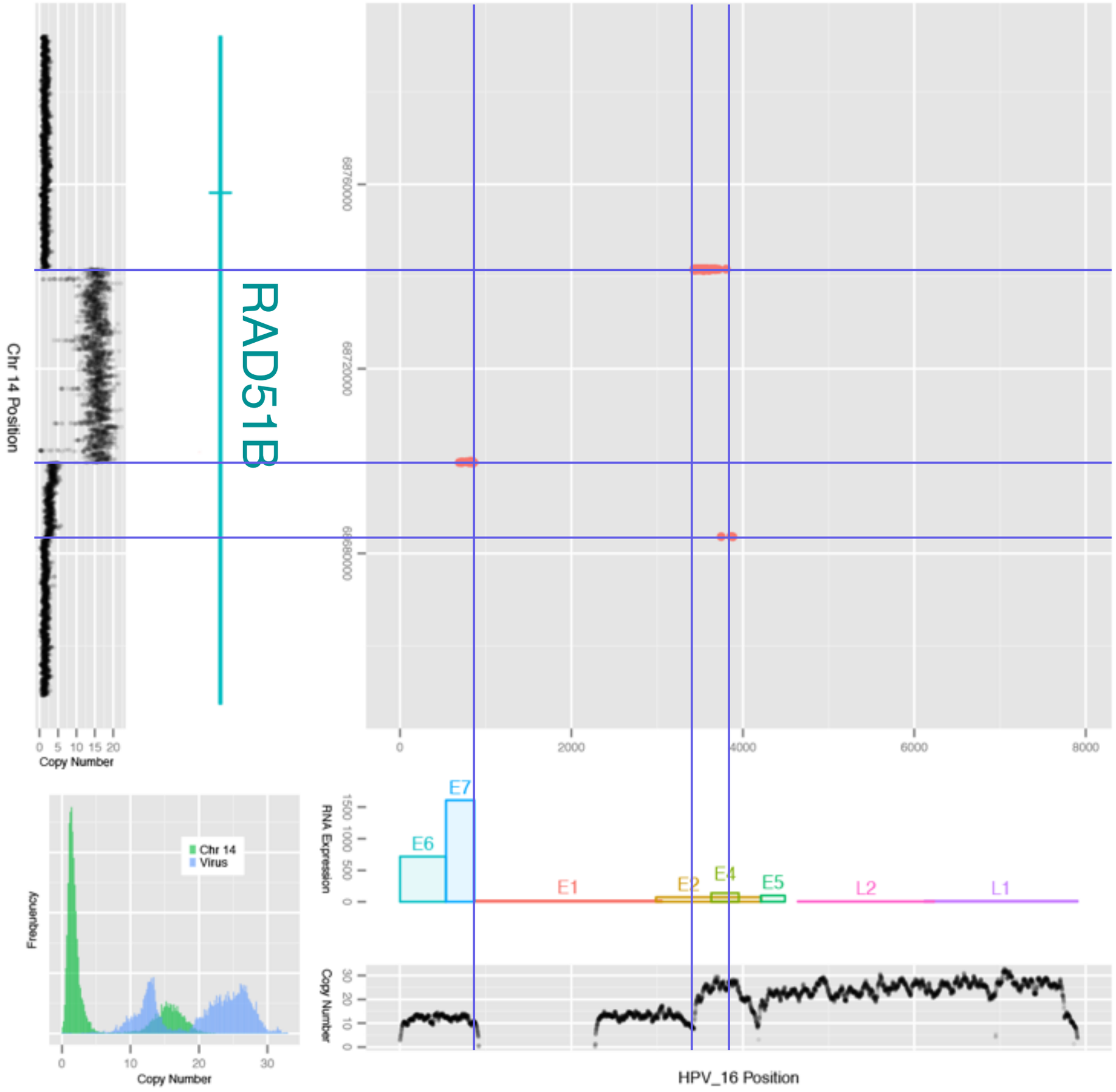
118 Discordant Pairs



Head and Neck

BA-4077-01

118 Discordant Pairs



RAD51B Exon Expression

RNA Expression Level (RSEM)

10.0

0.1

RAD51B Exon

BA–4077

HNSC tumors

RAD51B Exon Expression

RNA Expression Level (RSEM)

10.0

0.1

Integration

5’

3’

RAD51B Exon

BA–4077

HNSC tumors

BA–4077 Exon Expression Near Virus Integration (+/– 1 Mbp)

$\ln \frac{RSEM}{\langle RSEM \rangle}$

2

0

−2

68000000

68500000

69000000

69500000

Chr 14 Position

Gene Strand

–

+

EIF2S1+

PLEKHH1+

RAD51B+

PLEK2–

VTI1B–

ZFYVE26–

PPIAP6–

ZFP36L1–

BANF1P1+

DDX18P1+

ATP6V1D–

TMEM229B–

ARG2+

RDH12+

ACTN1–AS1+

GALNTL1+

MPP5+

PIGH–

RDH11–

ACTN1–

DCAF5–

EXD2+

Gene expression near integration (100 Kbp, CNV +)

$\ln \frac{RSEM}{\langle RSEM \rangle}$

4
2
0
−2
−4

Upstream (5’)
Within Integration
Downstream (3’)

Gene expression near integration (100 Kbp, CNV –)

$\ln \frac{RSEM}{\langle RSEM \rangle}$

6
4
2
0
−2
−4

Upstream (5')
Within Integration
Downstream (3')

Summary

- Developed a pipeline for multi-modal integration of TCGA data.
    - RNA-Seq for virus discovery and expression analysis,
    - WGS, WXS for integration analysis,
    - Unified visual representation of integration events.
- Illustrated close association between virus integration, copy number domains, and expression levels.

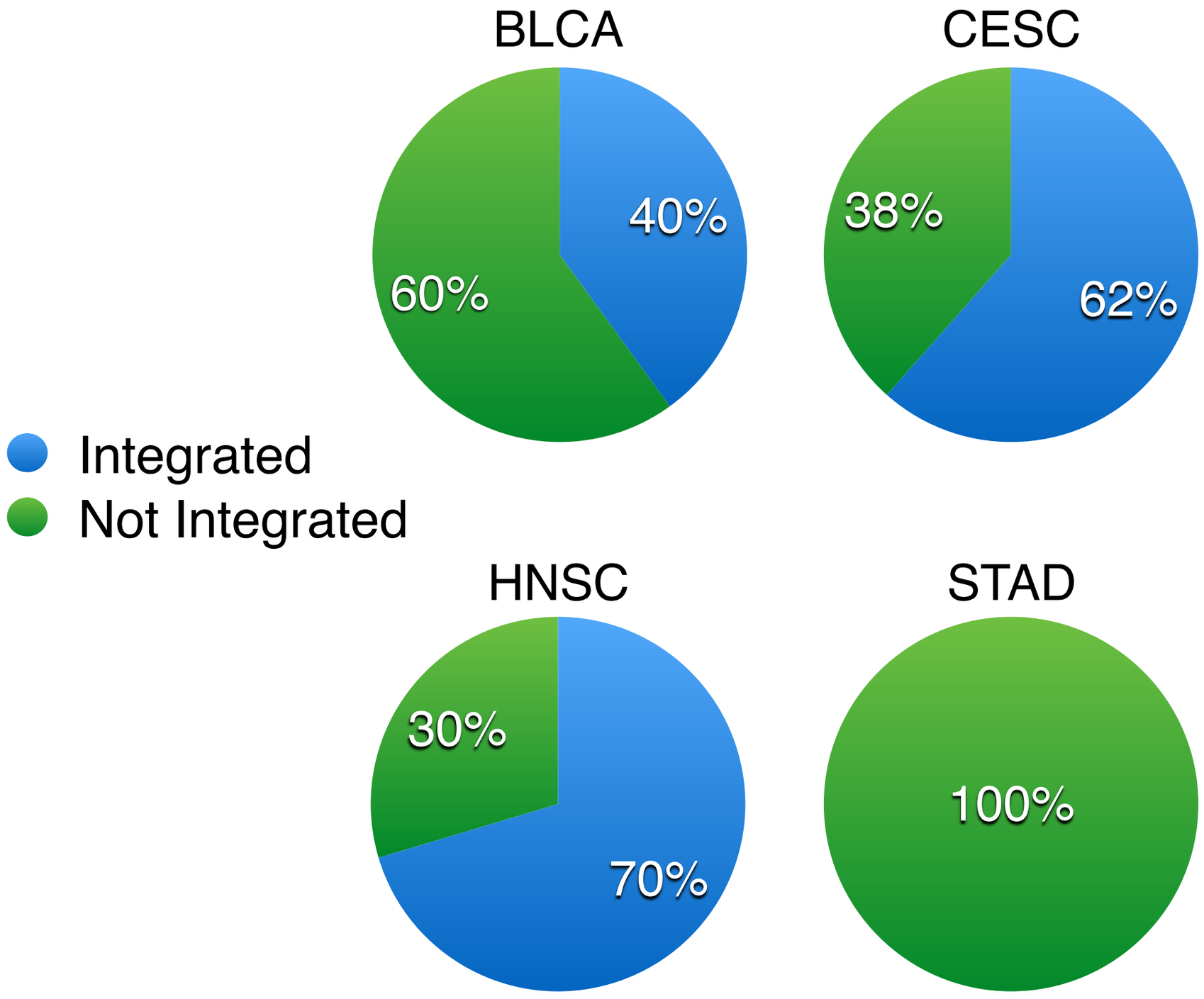
Acknowledgements

- **Li Ding**
- Kristine Wylie
- Song Cao
- Michael D. McLellan
- Manish Boolchandani
- Kai Ye
- Janet Rader
- George Weinstock

NATIONAL® CANCER INSTITUTE

National Human Genome Research Institute

Integration by Tumor



Select Virus Reference

- GRCh37 Human
- HPV6
- HPV16
- HPV18
- HPV31
- HPV33
- HPV35
- HPV39
- HPV45
- HPV52
- HPV56
- HPV58
- HPV59
- BK (Polyoma BK)
- HPyV7 (Polyoma HPyV7)
- HHV1
- EBV (HHV 4 / Epstein Barr)
- HHV5 (Cytomegalovirus)
- HBV (Hepatits B)
- Merkel Cell Polyoma