

# Intermediate *R* / *Bioconductor* for Sequence Analysis

Marc Carlson, Valerie Obenchain, Hervé Pagès, Paul Shannon, Dan Tenenbaum, Martin Morgan<sup>1</sup>

14-15 February 2013

<sup>1</sup>[mtmorgan@fhcrc.org](mailto:mtmorgan@fhcrc.org)

# Contents

<b>I</b>	<b><i>R / Bioconductor</i></b>	<b>5</b>
<b>1</b>	<b><i>R</i></b>	<b>6</b>
1.1	Statistical analysis and comprehension . . . . .	6
1.2	Essentials . . . . .	6
1.3	Do's . . . . .	6
1.4	Help! . . . . .	6
<b>2</b>	<b><i>Bioconductor</i></b>	<b>7</b>
2.1	Packages . . . . .	7
2.2	Classes and methods . . . . .	7
2.3	Help! . . . . .	7
<b>3</b>	<b>Sequencing</b>	<b>8</b>
3.1	Technologies . . . . .	8
3.2	Data . . . . .	8
3.3	Biology . . . . .	8
3.4	Analysis . . . . .	8
3.4.1	Work flows . . . . .	8
3.4.2	Who does what, and why? . . . . .	8
<b>4</b>	<b>Reads and strings</b>	<b>9</b>
4.1	Reads . . . . .	9
4.2	<i>Biostrings</i> . . . . .	9
4.3	<i>ShortRead</i> . . . . .	9
<b>5</b>	<b>Alignments and Ranges</b>	<b>10</b>
5.1	Alignments . . . . .	10
5.2	<i>GenomicRanges</i> . . . . .	10
<b>II</b>	<b>Differential Representation</b>	<b>11</b>
<b>6</b>	<b>RNA-seq Data</b>	<b>12</b>
6.1	Experimental considerations . . . . .	12
6.1.1	Differential expression . . . . .	12

6.1.2	Novel transcripts . . . . .	12
6.2	Work flows . . . . .	12
6.2.1	<i>Bioconductor</i> software . . . . .	12
6.2.2	Third-party software . . . . .	12
6.3	Example data: <i>pasilla</i> . . . . .	12
<b>7</b>	<b>Statistical Considerations</b>	<b>13</b>
7.1	Types of analysis . . . . .	13
7.1.1	Designed experiments . . . . .	13
7.1.2	Exploratory analysis . . . . .	13
7.2	Lessons from micro-arrays . . . . .	14
7.2.1	Replication . . . . .	14
7.2.2	Multiple comparisons . . . . .	14
7.2.3	Samples versus genes . . . . .	14
7.2.4	Bias and artifact . . . . .	14
7.3	Statistical challenges in sequence data . . . . .	14
7.3.1	Replication . . . . .	14
7.3.2	Counts . . . . .	14
7.3.3	Bias . . . . .	14
7.3.4	Technical artifacts . . . . .	14
<b>8</b>	<b><i>DESeq</i> Work Flow</b>	<b>15</b>
8.1	Data input and preparation . . . . .	15
8.2	Inference . . . . .	15
8.3	Independent filtering . . . . .	16
8.4	Data quality assessment . . . . .	16
8.4.1	Preliminary transformation . . . . .	16
8.4.2	Quality assessment . . . . .	16
8.5	Frequently asked questions . . . . .	16
<b>9</b>	<b>Additional Work Flows</b>	<b>18</b>
9.1	Additional work flows in <i>DESeq</i> . . . . .	18
9.1.1	Review . . . . .	18
9.1.2	More complicated designs . . . . .	18
9.2	<i>edgeR</i> . . . . .	18
9.2.1	Simple work flows . . . . .	18
9.2.2	More complicated designs . . . . .	18
9.3	Additional packages . . . . .	18
9.3.1	<i>BitSeq</i> . . . . .	18
9.3.2	<i>DEXSeq</i> . . . . .	18
<b>III</b>	<b>Variant Calls</b>	<b>19</b>
<b>10</b>	<b>Data for Variant Analysis</b>	<b>20</b>
10.1	DNA sequencing . . . . .	20

10.2	Work flows	20
10.2.1	<i>Bioconductor</i> software	20
10.2.2	Third-party software	20
10.3	Example data: lung cancer cell lines	20
<b>11</b>	<b><i>VariantTools</i> Work Flow</b>	<b>21</b>
11.1	Calling single-sample variants	21
11.1.1	Filters	21
11.1.2	Called variants	21
11.1.3	Data export	21
11.2	Additional work flows	21
11.2.1	Comparing variants across samples	21
11.2.2	Finding wild-type and no-call regions	21
<b>12</b>	<b>Working with Called Variants</b>	<b>22</b>
12.1	Data input with <i>VariantAnnotation</i>	22
12.1.1	Filtering and sorting	22
12.1.2	Input	22
12.2	Annotation	22
12.3	SNPs	22
<b>IV</b>	<b>Annotation and Visualization</b>	<b>23</b>
<b>13</b>	<b>Gene-centric Annotation</b>	<b>24</b>
13.1	Packages for genes, pathways, and model organisms	24
13.2	<i>biomaRt</i> and other web-based resources	24
13.3	Advanced aspects of <i>*.org</i> packages	24
13.3.1	<i>AnnotationForge</i> to create custom packages	24
13.3.2	<i>mysql</i> data base and SQL queries	24
<b>14</b>	<b>Genomic Annotation</b>	<b>25</b>
14.1	Whole genome sequences	25
14.1.1	<i>BSgenome.*</i> packages for model organisms	25
14.1.2	<i>FaFile</i> for simple whole-genome FASTA files	25
14.2	Gene models	25
14.2.1	<i>TxDb.*</i> packages for model organisms	25
14.2.2	Easily creating <i>TranscriptDb</i> objects from GTF files	25
14.3	UCSC tracks	25
14.3.1	Using <i>rtracklayer</i>	25
14.4	Other genome-scale annotations	25
<b>15</b>	<b>Visualizing Sequence Data</b>	<b>26</b>
15.1	<i>GViz</i>	26
15.2	<i>ggbio</i>	26
15.3	Additional opportunities	26

15.3.1	Base graphics: strategies for working with large data . . . . .	26
15.3.2	<i>shiny</i> for easy interactive reports . . . . .	26
16	<b>A Look into the Future: <i>AnnotationHub</i></b>	<b>27</b>
V	<b>Appendix</b>	<b>28</b>
	<b>References</b>	<b>29</b>