

## Introduction to RNA-Seq data analysis on HPRC

Shichen Wang, PhD

Bioinformatics Scientist

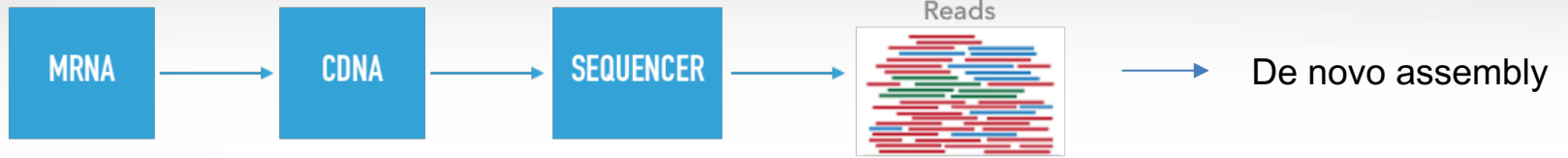
Genomics and Bioinformatics, AgriLife research

High Performance Research Computing



**DIVISION OF RESEARCH**  
TEXAS A & M UNIVERSITY

# RNA-Seq Overview



Assign reads to transcripts by aligning reads to reference genome or other alignment-free approaches



Count table

GENE NAME	TREAT_REP1	TREAT_REP2	TREAT_REP3	CONTROL_REP1	CONTROL_REP2	CONTROL_REP3
GENE0001	12	13	14	3	9	6
GENE0002	20	6	2	7	17	13
GENE0003	6	18	11	15	10	19
GENE0004	11	3	20	9	9	10
GENE0005	9	1	14	10	10	20
GENE0006	14	18	7	11	11	18
GENE0007	10	8	14	9	20	17
GENE0008	20	15	8	3	16	7

## WHAT DOES RNASEQ DATA PROVIDE US?

- ▶ Measure gene expression (relatively)
- ▶ Annotate transcripts
- ▶ Discover novel transcripts/isoforms
- ▶ Discover nucleotide variations

# RNASEQ EXPERIMENT DESIGN

- ▶ Questions often being asked:
  - ▶ How many replicates should I have?
  - ▶ How many reads should be generated?
- ▶ WHAT IS YOUR PRIMARY EXPERIMENTAL OBJECTIVE?
  - ▶ Detect DEGs
  - ▶ Annotation transcripts
  - ▶ Detect nucleotide variations



# Biological vs Technical Replication

- Biological replicates include multiple samplings within a population
- Technical replicates include multiple prepping and or resequencing the same individual
- Biological replicates generally increase statistical power more than technical replicates
  - Biological variability is generally greater than technical variability
  - Biological replicates contain both biological and technical variability





**Table 1.1 Recommendations for RNA-seq options based upon experimental objectives.**

Criteria	Annotation	Differential Gene Expression
Biological replicates	Not necessary but can be useful	Essential
Coverage across the transcript	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not as important; however the only reads that can be used are those that are uniquely mappable.
Depth of sequencing	High enough to maximize coverage of rare transcripts and transcriptional isoforms	High enough to infer accurate statistics
Role of sequencing depth	Obtain reads that overlap along the length of the transcript	Get enough counts of each transcript such that statistical inferences can be made
DSN	Useful for removing abundant transcripts so that more reads come from rarer transcripts	Not recommended since it can skew counts
Stranded library prep	Important for de Novo transcript assembly and identifying true anti-sense transcripts	Not generally required especially if there is a reference genome
Long reads (>80 bp)	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not generally required especially if there is a reference genome
Paired-end reads	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not important

<http://maseq.uoregon.edu>



# How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

Nicholas J. Schurch<sup>1,6</sup>, Pietá Schofield<sup>1,2,6</sup>, Marek Gierliński<sup>1,2,6</sup>,  
Christian Cole<sup>1,6</sup>, Alexander Sherstnev<sup>1,6</sup>, Vijender Singh<sup>2</sup>, Nicola Wrobel<sup>3</sup>,  
Karim Gharbi<sup>3</sup>, Gordon G. Simpson<sup>4</sup>, Tom Owen-Hughes<sup>2</sup>, Mark Blaxter<sup>3</sup> and  
Geoffrey J. Barton<sup>1,2,5</sup>

Author Affiliations

Corresponding authors: [g.g.simpson@dundee.ac.uk](mailto:g.g.simpson@dundee.ac.uk), [t.a.owenhughes@dundee.ac.uk](mailto:t.a.owenhughes@dundee.ac.uk),  
[Mark.Blaxter@ed.ac.uk](mailto:Mark.Blaxter@ed.ac.uk), [g.j.barton@dundee.ac.uk](mailto:g.j.barton@dundee.ac.uk)

←<sup>6</sup> These authors contributed equally to this work.

# Data analysis on HPRC system



# Where to Find NGS Tools

- TAMU HPRC Documentation
  - <https://hprc.tamu.edu/wiki/index.php/Ada:Bioinformatics>
- Type the following UNIX **commands** to see which tools are already installed on Ada
  - `module avail`
  - `module spider toolname` (not case sensitive, but read entire output)
  - `module key assembly` (some modules may be missed because this searches tool descriptions)
- If you find a tool that you want installed on Ada, send an email with the URL link to: `help@hprc.tamu.edu`
  - SeqAnswers <http://seqanswers.com/wiki/Software/list>
  - omictools.com
  - slideshare.net – find shared NGS presentations

# Finding NGS job template scripts using GCATemplates on Ada

```
mkdir $SCRATCH/rnaseq_class
```

```
cd $SCRATCH/rnaseq_class
```

```
module load GCATemplates
```

```
gcatemplates
```

For practice, we will copy a template file

- Select #13 RNA-seq, #1 QC, #1 rnaseqc, #1 two samples
- Final step will save a template job script file to your current working directory
- After you save the template file:

```
module purge
```

## Genomic Computational Analysis Templates

```
3*OTNFORMATTCS GCATemplates (ada)

CATEGORY
1. BAM files
2. ChIP-seq
3. FASTA files
4. FASTQ files
5. Functional genomics
6. Genome assembly
7. Genotyping
8. Metagenomics
9. Oxford Nanopore tools
10. PacBio tools
11. Phylogenetics
12. Population genetics
13. RNA-seq
14. SNPs & indels
15. Sequence alignments
16. Simulate data

s search
q quit

Select: 4
```



# QC Evaluation

```
module spider fastqc
```

- Use FastQC to visualize quality scores
  - Displays quality score distribution of reads
    - Input is a fastq file or files
    - Can disable grouping of sequence regions
  - Will alert you of poor read characteristics
  - Displays a representative sample of the fastq file
  - Can be run as a GUI or a command line interface
- FastQC will process using one CPU core per file
  - If there are 10 fastq files to analyze and 4 cores used
    - 4 files will start processing and 6 will wait in a queue
    - If there is only one fastq file to process then using 10 cores does not speed up the process

# QC Quality Trimming

- Sequence quality trimming tools

```
module spider Trimmomatic
```

← recommended tool

- Trimmomatic will maintain paired end read pairing after trimming
- Trim reads based on quality scores
  - Trim the same number of bases from each read or
  - Use a sliding window to calculate average quality at ends of sequences
- Decide if you want to discard reads with Ns
  - some assemblers replace Ns with As or a random base G, C, A or T
- Trim adapter sequences
  - Trimmomatic has a file of Illumina adapter sequences

```
module load Trimmomatic/0.36-Java-1.8.0_92
```

```
ls $EBROOTTRIMMOMATIC/adapters/
```



# RNA-SeQC

module spider RNA-SeQC

- Provides alignment metrics & graphs all samples together
  - Yield alignment and duplication rates
  - GC bias
  - rRNA content
  - Regions of alignment (exon, intron, intragenic)
  - continuity of coverage
  - 5'/3' bias and much more ...
- Metrics can help identify sample outliers by comparing metrics of all samples

## **RNA-SeQC: RNA-seq metrics for quality control and process optimization**

DeLuca, et al. [Bioinformatics](#). 2012 Jun 1; 28(11): 1530–1532. Published online 2012 Apr 25. doi: [10.1093/bioinformatics/bts196](#) PMID: PMC3356847

# Mapping RNA-seq Reads to a Reference Assembly

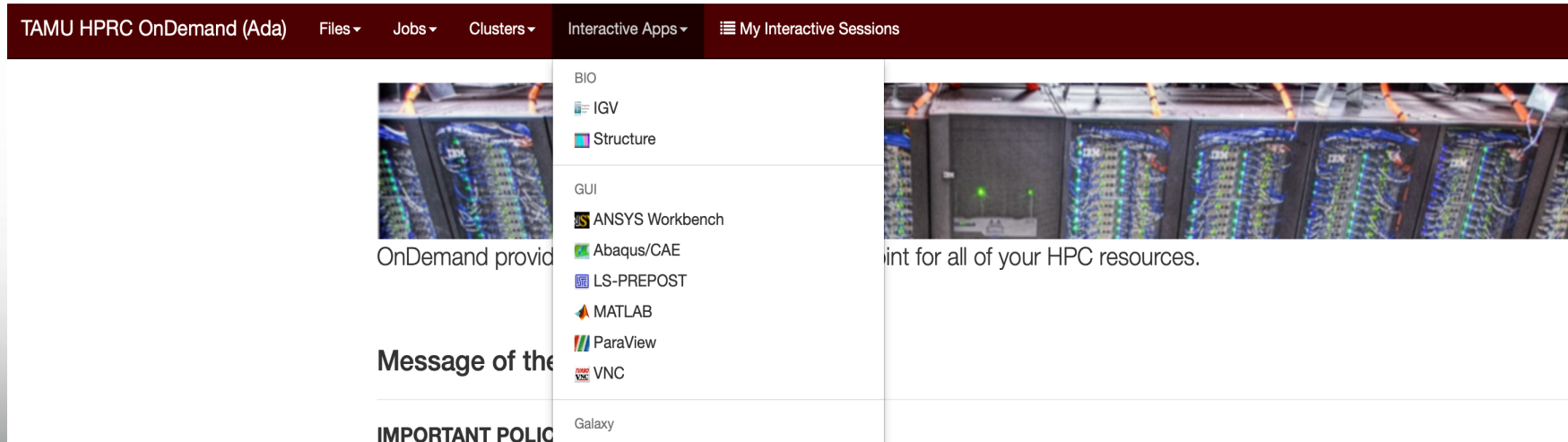


# Splice-Aware Aligners for RNA-seq Short Reads

- HISAT2 which supersedes TopHat2
  - `/scratch/datasets/genome_indexes/ucsc/mm10/hisat2_index/`
- STAR (on Ada as module STAR-STAR)
  - Uses gene annotations in gtf format
    - can use `gffread` in Cufflinks module to convert gff3 to gtf
  - supports PacBio but should use non-default settings
    - *Bioinfx study: Optimizing STAR aligner for Iso Seq data*
- BMap
  - also supports PacBio and Nanopore
- GMap
  - also supports PacBio and Nanopore

# Integrative Genomics Viewer (IGV) Exercise

- IGV is a genome browser with pre-loaded genomes available in which you can use to view multiple .bed, .sam and .vcf files.
- Running IGV through **portal.hprc.tamu.edu**



The screenshot shows the TAMU HPRC OnDemand portal interface. The top navigation bar includes "TAMU HPRC OnDemand (Ada)", "Files", "Jobs", "Clusters", "Interactive Apps", and "My Interactive Sessions". The "Interactive Apps" dropdown menu is open, displaying a list of applications categorized by type:

- BIO**
  - IGV
  - Structure
- GUI**
  - ANSYS Workbench
  - Abaqus/CAE
  - LS-PREPOST
  - MATLAB
  - ParaView
  - VNC
- Galaxy**

Below the application list, there are several text elements: "OnDemand provid", "Message of the", and "IMPORTANT POLIC". To the right of the application list, there is a large image of server racks with the text "point for all of your HPC resources."

# IGV viewing indexed bam file

IGV (on login7)

File Genomes View Tracks Regions Tools GenomeSpace Help

Mouse (mm10) chr11 [sparc] Go

Type sparc then click the Go button

Right click in this area and select "View as pairs"

Right click and select "Expanded"

chr11:55,408,131

507M of 1.171M

# RNA-seq for Differential Expression





# RNA-seq Sequence Fragment Counting

- Alignment based
  - Non-normalized alignment counts
    - HTSeq-count
  - Normalized (RPKM, FPKM, TPM)
    - eXpress (outputs FPKM)
    - RSEM (isoform/gene level estimates without RPKM or FPKM)
    - Trinity Transcript Quantification
      - A Trinity script can run: Kallisto, RSEM, eXpress, Salmon
- Non-Alignment based
  - Kallisto (pseudoalignment)
  - Salmon (lightweight alignment)
  - Sailfish (k-mer)

# RPKM vs FPKM vs TPM

- The number of **R**eads **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for single end reads
- The number of **F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for paired-end reads
    - If both paired reads align to a transcript then they are counted as one alignment
- **T**ranscripts **P**er kilobase **M**illion
  - Normalize for gene length first
  - Normalize for sequence depth second

<http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/>





# Tuxedo Suite

- HISAT2
  - splice aware mapping of RNA-seq reads
  - TopHat (which uses Bowtie2) and HISAT are superseded by HISAT2
- Cufflinks
  - assembles aligned reads into transcripts and estimates their abundances
- Cuffdiff
  - compares RNA-seq abundance (expression) levels of two samples or groups

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
CAWT_00001	CAWG_00001	-	chr_1.1:8373-9093	q1	q2	OK	111.944	163.869	0.549763	0.768107	0.58795	0.996768	no
CAWT_00002	CAWG_00002	-	chr_1.1:11447-12425	q1	q2	OK	14.5992	30.9037	1.08189	1.3841	0.2921	0.98312	no
CAWT_00003	CAWG_00003	-	chr_1.1:14130-14451	q1	q2	OK	248.323	259.152	0.0615814	0.172186	0.94685	0.996768	no
CAWT_00004	CAWG_00004	-	chr_1.1:14890-16045	q1	q2	OK	60.9546	86.0009	0.496617	0.604904	0.6204	0.996768	no
...													
CAWT_01628	CAWG_01628	-	chr1.2:664522-665344	q1	q2	OK	3.56447	157.849	5.46871	6.64693	0.00015	0.0482417	yes

p\_value = The uncorrected p-value of the test statistic.

q\_value = The FDR-adjusted p-value of the test statistic

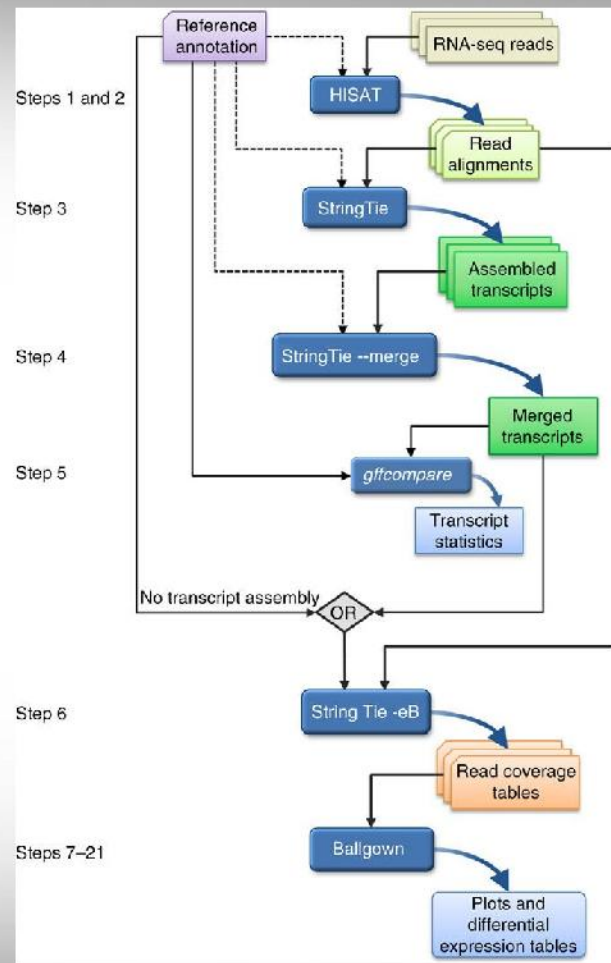


# “New Tuxedo” Protocol

## Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Pertea, et al. Nature Protocols 11,1650–1667 (2016)  
doi:10.1038/nprot.2016.095

HISAT2 supercedes HISAT



# Sailfish

- Alignment-free isoform quantification from RNA-seq data (uses k-mers)
- Requires a set of target transcripts (fasta)
  - From a reference or a *de novo* assembly
- Requires sequence reads (fasta or fastq)

Name	Length	EffectiveLength	TPM	NumReads
TRINITY_DN30_c0_g1_i1	215	68.4635	236.773	233
TRINITY_DN43_c0_g1_i1	280	102.34	5971.5	8784
TRINITY_DN88_c0_g1_i1	217	69.3036	191.74	191
TRINITY_DN59_c0_g1_i1	393	194.337	4092.64	11432
TRINITY_DN98_c0_g1_i1	205	64.4299	1097.09	1016
TRINITY_DN17_c0_g1_i1	310	122.99	2634.35	4657

# R Bioconductor

- Popular R bioconductor packages for RNA-seq
  - CQN – Normalization of RNA-seq data
  - edgeR – Differential gene expression
  - DESeq, DESeq2 – Differential gene expression
  - cummeRbund – analysis/visualization of cufflinks data
- Bioconductor packages can be found in this R version

```
module load R_tamu/3.3.1-intel-2015B-default-mt
```

# RNA-seq for Transcriptome Assembly





# RNA-seq Transcriptome Assembly

- Assembly with a reference genome

```
module spider Trinity
```

```
module spider HISAT2 Cufflinks
```

```
module spider Scripture
```

```
module spider StringTie
```

- *de novo* assembly without a reference genome

```
module spider Trinity
```

```
module spider Oases
```

# Running Trinity on Ada

- Trinity uses 100,000s of intermediate files
  - Contact **help@hprc.tamu.edu** and request a file quota increase before running Trinity
  - Run one Trinity job at a time and check resource usage
    - `showquota`
    - It is recommended not to run multiple Trinity jobs unless you know memory usage and an estimate of the number of temporary files
  - Trinity creates checkpoints and can be restarted if it stops due to file/disk quota met, out of memory or runtime
    - Checkpoints are not available when running Trinity in Galaxy
    - Checkpoints are not available if you use \$TMPDIR with Trinity
      - need to rsync results from \$TMPDIR at end of job script
      - checkpoints are stored in \$TMPDIR which is deleted after job ends
- See GCATemplates for sample Trinity scripts

# Transcriptome Assembly Completeness

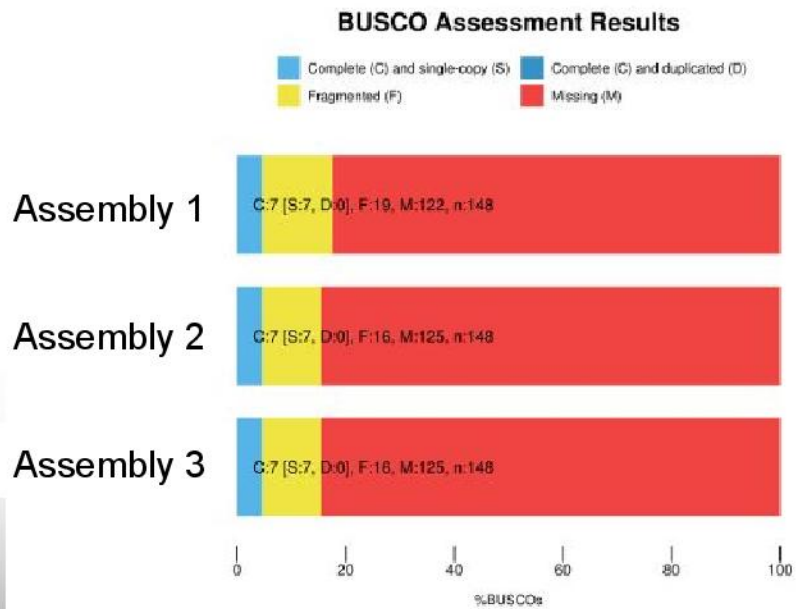
The completeness of a transcriptome can be estimated by using a set of highly conserved genes that are common to specific taxonomic groups

- 44 taxonomic groups available
  - aves, bacteria, eukaryota, insecta, vertebrata, ...
- BUSCO – uses single-copy genes to assess transcriptome assembly and annotation completeness
  - evaluates % complete 'BUSCOs', % fragmented, % missing
  - can run in genome, transcriptome or protein mode
  - `module spider BUSCO`



# Transcriptome Assembly Completeness

BUSCO script (`generate_plot.py`) can be used to plot multiple BUSCO short summaries to compare different assemblies



Any questions?



# DEG Analysis tutorial:

1. Quality check: FastQC
2. Trim adapters and low quality bases: Trimmomatic
3. Alignment to the reference: HISAT2
4. Reference guided assembly: cufflinks
5. Merge the transcripts: cuffmerge
6. Differential expression analysis: edgeR
7. Visualize the results: heatmap and volcano plot

## Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

### Use Galaxy



Use project's free server or other public servers

### Get Galaxy



Install locally or in the cloud or get Galaxy on SlipStream

### Learn Galaxy



Screencasts, Galaxy 101, ...

### Get Involved



Mailing lists, Tool Shed, wiki



## TEXAS A&M HIGH PERFORMANCE RESEARCH COMPUTING

Home

User Services

Resources

Research

Policies

Events

About

<https://galaxy-terra.hprc.tamu.edu/bdf>

Thanks to Dr. Charles Michael Dickens!



Workflow Visualize Shared Data Admin Help User Galaxy

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494