## Appendix A

# Bioinformatics workflow

Table A.1: `Velvet` assembly statistics of contigs longer than 1 000 bp for a single lane of paired 76 bp sequences from *Eucalytpus* xylem tissue reads trimmed to different lengths. The assemblies were all performed with a kmer setting of 41. These statistics were used to calculate the assembly score, as discussed in Section 2.3.3 on page 56 and presented in Table 2.6.

| Read length | N | Sum | Min | 1st Quartile | Median | 3rd Quartile | Max | Mean | N50 |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 2 644 | 3 853 938 | 1 000 | 1 118 | 1 300 | 1 611.5 | 6 772 | 1 457.61 | 1 424 |
| 55 | 5 045 | 7 722 735 | 1 000 | 1 138 | 1 342 | 1 709 | 8 078 | 1 530.77 | 1 512 |
| 60 | 6 458 | 10 216 572 | 1 000 | 1 149 | 1 371 | 1 770 | 8 241 | 1 582.00 | 1 574 |
| 65 | 7 165 | 11 547 759 | 1 000 | 1 160.5 | 1 393 | 1 804 | 11 049 | 1 611.69 | 1 609 |
| 70 | 7 548 | 12 288 379 | 1 000 | 1 162 | 1 395 | 1 823 | 11 008 | 1 628.03 | 1 627 |
| 76 | 7 857 | 12 917 451 | 1 000 | 1 164 | 1 415 | 1 848 | 9 925 | 1 644.06 | 1 643 |

## Appendix B

# Extendinator

The `Python` script used for a coverage-assisted re-assembly of contigs, also known as ”extendinator” is provided on the following pages. A graphical representation of the process is provided in Figure 3.1. The program selects an entry from the assembled contigs file, and performs an alignment of the short reads to the selected contig and calculated the true coverage of the contig. After alignment, the program extracts all the short reads together with their respective mate-pairs from a `Berkeley` database, and sends the contig as well as the sampled short reads to `Velvet` with the calculated coverage parameter to perform a directed contig assembly.

```
"""
    Extendinator:
        An interative approach to try and improve contig sizes.

    1) Map all the short reads to a contig, get the reads that mapped.
    2) Extract the pairs
        2a) Connect to a database, get all the reads that match
            2ab) Convert to fasta
    3) Assemble with Velvet
        3a) Parameter range cc_9 ec [9,50,100,200,1000]
        3b)Join the longest assemblies in one file (best_assembly.fa)


    @requires: Biopython
    @requires: bsddb3
    @author: charles.hefer@gmail.com
"""


import sys
import getopt
from datetime import datetime
from Bio import SeqIO
import os
import subprocess
import time
from multiprocessing import Process
import bsddb3

global usage
usage = """
Extendinator: An iterative approach to extext Velvet contigs

Usage: python start_extendinator.py [options] short_reads.fa contigs.fa

++Bowtie options++
\t-f\t--short_reads_type\Either fa for fasta, or fq for fastq, default is fa
\t-b\t--bowtie_mismatch\tNumber of mismatches allowed during the bowtie matching of the short reads to the contig
\t-m\t--max_bowtie_processes\tMax number of bowtie processes
\t-t\t--threads\tNumber of threads for Bowtie, this times the #processes = number of CPUs

++Global options++
\t-h\t\t--help\tThis help message
"""
global cwd
cwd = os.getcwd()
global bowtie_build_cmd
bowtie_build_cmd = "/usr/local/bowtie/bowtie-build"
global bowtie_cmd
bowtie_cmd = "/usr/local/bowtie/bowtie"
global bdb
bdb = "./pairs.db"


class UsageEx(Exception):
    """The standard exception"""
    def __init__(self, msg):
        """
            Sets the exception message
            @var msg: The exception message thrown
        """
        self.msg = msg

def now():
    """
        Converts the current time to a string format

        @requires: datetime.datetime
        @return: A string reprepsentation of datetime.now()
    """
    curr_time = datetime.now()
    return curr_time.strftime("%c")


def get_number_of_processes(process):
    """Returns the number of processes returned by grep
    ps -eaf | grep processname
    Subtract the grep itself, and the extra newline that comes through.
    @var process: The process to grep for
    @type process: String

    @return: The number of process as an int
    """
    num_procs = subprocess.Popen("ps -eaf | grep '%s'" % process, shell=True, stdout=subprocess.PIPE)
    output = num_procs.stdout.readlines()
    i = len(output) - 2
    return i
```

```
def multiprocess_start(cmd):
    """
        Executes the command as a multiprocess
    """
    process = subprocess.call(cmd, shell=True, stdout=subprocess.PIPE)
    return process


def prepare_bowtie_build(dir, filename, max_bowtie_processes):
    """
        Sets the command to run bowtie build on the contig
    """
    #the resulting build has a _ewbt extention
    #and is in the ./bowtie dir
    cmd = "%s %s %s_ewbt" % (bowtie_build_cmd, dir+filename, "./bowtie/"+filename)

    while get_number_of_processes("bowtie_build") >= max_bowtie_processes:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()


def prepare_bowtie_align(short_reads_filename, ewbt_filename, bowtie_mismatch, max_bowtie_processes, threads, short_reads_filetype):
    """
        Aligns the short reads to the file
    """
    cmd = "%s -%s -n %s --alfa=%s.match -p %s %s %s %s.out" % (bowtie_cmd,
                                                                short_reads_filetype,
                                                              bowtie_mismatch,
                                                              "bowtie/"+ewbt_filename,
                                                              threads,
                                                              "bowtie/"+ewbt_filename,
                                                              short_reads_filename,
                                                              "bowtie/"+ewbt_filename)

    while get_number_of_processes("bowtie") >= max_bowtie_processes:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()

def save_biopython_entry(dir, entry, format):
    """
        Saves the biopython object in the correct format
    """
    try:
        handle = open(dir+"/" + entry.name + "/" + entry.name+".fa", "w")
    except IOError, e:
        print(e)
        sys.exit()
    SeqIO.write([entry], handle, format)
    handle.close()
    return dir+"/"+entry.name + "/" + entry.name + ".fa"

def bowtie_watcher(contig, max_bowtie_processes, bowtie_mismatch,short_reads_filename):
    """
        Somehow manages the number of bowtie executables that can be started
    """
    #Get the current number of bowties running
    current = get_number_of_processes("bowtie")
    while current > max_bowtie_processes:
        time.sleep(10)
    else:
        bowtie_dir = prepare_bowtie_dir(contig.name)
        contig_file_name = save_biopython_entry(bowtie_dir, contig, "fasta")
        bowtie_builder(contig_file_name)
        bowtie_aligner(contig_file_name, bowtie_mismatch, short_reads_filename)

def split_fasta_file(handle, dir):
    """
        Takes every entry, create an output file for that entry in the dir
    """
    entries = SeqIO.parse(handle, "fasta")
    for entry in entries:
        out = open(dir+entry.name.replace(" ","").replace("\\","").replace("|","_").replace("/","_").replace("(","_").replace(")","_")
        SeqIO.write([entry], out, "fasta")
        out.close()

def create_mates_file(base_name, database_name):
    """ Iterates over ./bowtie/base_name.match, and returns all the mated
        that is found in the berkeley database
        Creates a file basename.fa in ./mates
    """

    try:
        handle = open("./bowtie/%s.fa_ewbt.match" % (base_name),"r")
        out_handle = open("./mates/%s.fa" % (base_name),"w")
    except IOError, e:
        #No alignments found... can do nothing about that
        #should this be reported?
```

```
        return None
    entries = SeqIO.parse(handle, "fasta")

    mate_pairs = []
    pairs = bsddb3.hashopen(bdb, "r")
    for entry in entries:
        out_handle.write(">%s\n" % entry.name)
        out_handle.write("%s\n" % pairs[entry.name].split(",")[0])
        out_handle.write(">%s\n" % entry.name)
        out_handle.write("%s\n" % pairs[entry.name].split(",")[1])
    out_handle.close()

def faLen_stats(file):
    """
        Returns the result from running faLen on the file
        #TODO: Rewrite use subprocess
    """
    import popen2

    output = []

    cmd = "faLen < %s | stats" % (file)
    process = popen2.Popen3(cmd)
    process.wait()
    result = process.fromchild.readlines()
    for line in result:
        line = line.replace(" ","")
        output.append(line.split("=")[1].rstrip())
    output.append("\n")
    return output


def velveth_runner(filename, kmer):
    """
        Runs velveth on the file, hashing for the kmer
    """
    velvet_exe = "/usr/local/velvet/velveth"
    cmd = "%s ./velvet/%s/assembly %s -fasta -shortPaired ./mates/%s -long ./fasta/%s" % \
    (velvet_exe, filename, kmer, filename, filename)

    while get_number_of_processes("velveth") >= 20:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()
    time.sleep(2)

def get_coverage(filename):
    """
        Returns the coverage value stored in ./mates/cov_stats.csv
    """
    file = open("./mates/cov_stats.csv", "r")
    for line in file:
        if filename in line:
            cols = line.split(",")
            contig_length = int(cols[1])
            bases = int(cols[2].rstrip())
    return bases/float(contig_length)


def velvetg_runner(filename):
    """
        Runs velvetg in the file, hashing for the kmer
    """
    velvet_exe = "/usr/local/velvet/velvetg"
    coverage = get_coverage(filename)

    cmd = "%s ./velvet/%s/assembly -ins_length 200 -ins_length_sd 80 -exp_cov %s -cov_cutoff 8" % \
    (velvet_exe, filename, coverage)

    print cmd

    while get_number_of_processes("velveth") >= 20:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()

def save_longest_entry(entry_name, contigs_file, location):
    """
        Finds the longest entry in the contigs_file, rename it to the
        entry name [minus the extention] , and saves it in the locatoion
    """
    try:
        contigs_handle = open(contigs_file, "r")
        location_handle = open(location+"/%s" % entry_name, "w")
    except IOError, e:
        print(e)

    longest_entry = None
    longest_length = 0
```

```
    entries = SeqIO.parse(contigs_handle, "fasta")
    for entry in entries:
        if len(entry.seq) > longest_entry:
            longest_entry = entry
            longest_length = len(entry.seq)

    #Rename the longest_entry
    longest_entry.id = entry_name.replace(".fa","")
    longest_entry.name = ""
    longest_entry.description = ""
    #write to the location
    SeqIO.write([longest_entry], location_handle, "fasta")
    location_handle.close()
    contigs_handle.close()

    #update the report
    #remove the entries that did not grow for fasta
    #repeat = True

def main(argv = None):
    """
        The main program flow
    """
    print("%s Extendinator started" % now())

    #Get all the arguments
    if argv is None:
        argv = sys.argv
    try:
        try:
            opts, args = getopt.getopt(argv[1:], "b:h:m:t:f:",
                                        ["bowtie_mismatch=",
                                        "max_bowtie_processes=",
                                        "threads=",
                                        "short_reads_type"
                                        "help"])

            bowtie_mismatch = 2
            max_bowtie_processes = 1
            threads = "2"
            short_reads_filetype = "f"


            for opt, value in opts:
                if opt in ("-b", "--bowtie_mismatch"):
                    bowtie_mismatch = value
                if opt in ("-m","--max_bowtie_processes"):
                    max_bowtie_processes = int(value)
                if opt in ("-t", "--threads"):
                    threads = value
                if opt in ("-f", "--short_reads_type"):
                    if value == "fq":
                        short_reads_filetype = "q"
                if opt in ("-h", "--help"):
                    print(usage)
                    raise sys.exit()


        except getopt.error, e:
            print(e)
            raise UsageEx(e)

        #test the presence of the contigs and short read files
        try:
            print("%s Validating the short reads file: %s" % (now(), args[0]))
            short_reads_filename = cwd+"/"+args[0]
            short_reads_handle = open(args[0],"r")
            print("%s Validating the contigs file: %s" % (now(), args[1]))
            contigs_handle = open(args[1],"r")
        except IOError,e:
            print(e)
            raise UsageEx(e)

        #Prepare the directory structure
        #this can be made more intelligent
        try:
            os.system("rm -rf bowtie")
            os.system("rm -rf fasta")
            os.system("rm -rf mates")
            os.system("rm -rf velvet")
        except OSError:
            pass
        try:
            os.mkdir("bowtie")
            os.mkdir("fasta")
            os.mkdir("mates")
            os.mkdir("velvet")
        except OSError, e:
            os.system("rm -rf bowtie/*")
```

```
    os.system("rm -rf fasta/*")
    os.system("rm -rf mates/*")
    os.system("rm - rf velvet/*")

try:
    report_handle = open("report.csv","w")
except IOError, e:
    print(e)
    sys.exit()

#The step is to parse the contigs file
print("%s Parsing the contigs file into ./fasta" % now())
split_fasta_file(contigs_handle, "./fasta/")
contigs_handle.close()

fasta_entries = os.listdir("./fasta")

#generate an file with the initial lengths
print("%s Generate the initial report template" % now())
report_handle.write("Sequence_entry,init_length\n")
for fasta_file in fasta_entries:
    #get the sequence length
    entry_length = int(faLen_stats("./fasta/%s" % fasta_file)[1])
    report_handle.write("%s,%s\n" % (fasta_file, entry_length))
report_handle.close()

while 1:
    fasta_entries = os.listdir("./fasta")

    if len(fasta_entries) == 0:
        break

    print("%s Building the Bowtie indices" % now())
    for fasta_entry in fasta_entries:
        prepare_bowtie_build("./fasta/", fasta_entry, max_bowtie_processes)
    time.sleep(2)
    #Need to wait for all the processes to finish
    while get_number_of_processes("bowtie_build") > 0:
        time.sleep(5)

    print("%s Running Bowtie aligner with %s mismatches" % (now(), bowtie_mismatch))
    print("Stdout from Bowtie to follow...this can be ignored")
    for fasta_entry in fasta_entries:
        prepare_bowtie_align(short_reads_filename, fasta_entry+"_ewbt", bowtie_mismatch, max_bowtie_processes, threads, short_
        #give the os time to register
        time.sleep(2)
    #Need to wait for all the processes to finish
    time.sleep(5)
    while get_number_of_processes("bowtie") > 0:
        time.sleep(5)
    print("%s Done with the Bowtie aligner" % (now()))


    print("%s Preparing to find the mates" % (now()))
    for fasta_entry in fasta_entries:
        #change the name
        fasta_entry = ".".join(fasta_entry.split(".")[:-1])
        create_mates_file(fasta_entry, bdb)
    time.sleep(5)
    print("%s Mates now in ./mates" % now())

    print("%s Calculating the coverage statistics" % now())
    mate_entries = os.listdir("mates")
    try:
        mate_entries.remove("cov_stats.csv")
    except:
        pass
    cov_stats_handle = open("mates/cov_stats.csv", "w")
    cov_stats_handle.write("Contig_name,Lenght,Bases_in_mates")
    cov_stats_handle.write("\n")
    for mate_entry in mate_entries:
        contig_length = int(faLen_stats("./fasta/%s" % mate_entry)[1])
        pairs_bases = int(faLen_stats("./mates/%s" % mate_entry)[1])
        cov_stats_handle.write("%s,%s,%s" % (mate_entry, contig_length, pairs_bases))
        cov_stats_handle.write("\n")
    cov_stats_handle.close()
    time.sleep(5)
    print("%s Finished with the coverage statistics, in ./mates/cov_stats.csv" % now())

    print("%s Preparing for the velvet hashing " % now())
    for entry in mate_entries:
        try:
            os.mkdir("./velvet/%s" % entry)
        except OSError, e:
            pass
        velveth_runner(entry,"31")
    time.sleep(5)
    while get_number_of_processes("velveth") > 0:
        time.sleep(5)
```

```
            print("%s Done with the velvet hashing" % now())

            print("%s Preparing for the velvet assembly " % now())
            for entry in mate_entries:
                velvetg_runner(entry)
            time.sleep(5)
            while get_number_of_processes("velveth") > 0:
                time.sleep(5)
            print("%s Done with the velvet assembly" % now())

            print("%s Getting the longest entry for every assembly" % now())
            for entry in mate_entries:
                #the contigs resides in velvet/entry/assembly/contigs.fa
                save_longest_entry(entry, "velvet/%s/assembly/contigs.fa" % entry, "fasta/")
            print("%s All the longest entries now back in ./fasta" % now())

            print("%s Adding the newest data to the report.csv file" % now())
            #Append to the reports file
            os.system("mv ./report.csv ./report.csv.prev")
            reports_handle = open("report.csv.prev","r")
            report_out_handle = open("report.csv","w")
            report_out_handle.write("Sequence_entry,init_length\n")
            for line in reports_handle:
                if line.startswith("Sequence_entry"):
                    continue
                line = line.rstrip()
                cols = line.split(",")
                #the name of the entry is the first col
                try:
                    entry_length = int(faLen_stats("./fasta/%s" % cols[0])[1])
                    cols.append("%i" % entry_length)
                except IndexError, e:
                    pass
                outline = ",".join(cols)
                report_out_handle.write(outline + "\n")
            report_out_handle.close()
            reports_handle.close()
            print("%s Updated the report.csv file" % now())

            #Now, check the report file, if the last entry is smaller or equal to
            #the second last entry, then call the entry finished
            #remove from ./fasta/
            #and append to finished_contigs.fa
            print("%s remove the contigs that does not want to grow any more" % now())
            report_handle = open("report.csv","r")
            for line in report_handle:
                if line.startswith("Sequence_entry"):
                    continue
                print line
                line = line.rstrip()
                cols = line.split(",")
                if int(cols[-1]) <= int(cols[-2]):
                    print cols[0]
                    os.system("less ./fasta/%s >> finished_contigs.fa" % cols[0])
                    os.system("rm ./fasta/%s" % cols[0])
                    os.system("rm ./mates/%s" % cols[0])
                    print os.listdir("fasta")
                    print os.listdir("mates")

            print("%s And start over again?" % now())
        print("%s Done" % now())


    except UsageEx, err:
        print(usage)

if __name__ == "__main__":
    if len(sys.argv) < 3:
        print(usage)
        sys.exit()
    else:
        sys.exit(main())
```

## Appendix C

# Transcriptome assembly

## C.1. Evaluating contig contiguity of the assembled transcript sequences

### C.1.1. Full length *Eucalyptus* cDNA sequences

The following table contains the 34 full length CDS sequences used to validate the assembly. The functional role of the 33 sequences ranges from transcription factors, transporter genes, structural and developmental proteins, indicating that the assembled transcriptome succesfully assembled near full length genes, including the 5' and 3' UTR regions for a wide variate of mRNA sequences.

| Accession | Contig_id | Description | length | FPKM |
|---|---|---|---|---|
| AB465730.1 | contig_87094 | Eucalyptus grandis AGL mRNA for agamous-like protein, complete cds. | 1184 | 17.98 |
| AB479542.1 | contig_10798 | Eucalyptus grandis mRNA for transcription factor Myb, complete cds. | 666 | 14.02 |
| AB479543.1 | contig_45922 | Eucalyptus grandis mRNA for transcription factor GRAS family protein, complete cds. | 1485 | 13.00 |
| AB479544.1 | contig_94920 | Eucalyptus grandis mRNA for 1-aminoacyclopropane-1-carboxylate oxidase, complete cds. | 1288 | 81.75 |

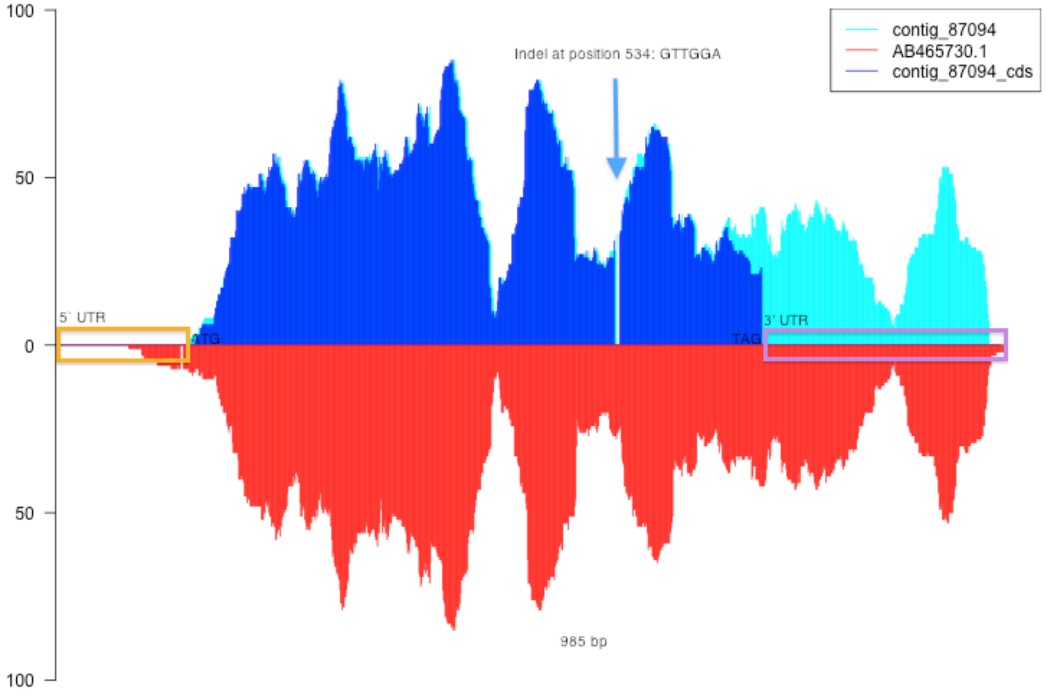
| | | | | |
|---|---|---|---|---|
| AB479545.1 | contig_56935 | Eucalyptus grandis mRNA for transcription factor squamosa promoter binding protein like, complete cds | 1940 | 43.35 |
| AF029976.1 | contig_93436 | Eucalyptus grandis MADS box protein (EGM2) mRNA, complete cds. | 920 | 13.01 |
| AF197329.1 | contig_5550 | Eucalyptus grandis zinc transporter (EgZnT1) mRNA, complete cds. | 1635 | 17.08 |
| AF197330.1 | contig_2649 | Eucalyptus grandis calcineurin-like protein (EgCBL1) mRNA, complete cds. | 951 | 27.21 |
| AY150283.1 | contig_11286 | Eucalyptus grandis fertilization independent endosperm development protein mRNA, complete cds | 1626 | 18.87 |
| AY263807.1 | contig_68957 | Eucalyptus grandis SOC1-like floral activator MADS3 mRNA, complete cds. | 1112 | 21.66 |
| AY263808.1 | contig_52396 | Eucalyptus grandis SOC1-like floral activator MADS4 mRNA, complete cds. | 980 | 8.80 |
| AY263809.1 | contig_6043 | Eucalyptus grandis SVP-like floral repressor mRNA, complete cds. | 855 | 20.09 |
| DQ014506.1 | contig_2805 | Eucalyptus grandis cellulose synthase 2 (CesA2) mRNA, complete cds. | 3471 | 226.37 |
| DQ014507.1 | contig_31 | Eucalyptus grandis cellulose synthase 3 (CesA3) mRNA, complete cds. | 3452 | 220.59 |
| DQ014509.1 | contig_4202 | Eucalyptus grandis cellulose synthase 5 (CesA5) mRNA, complete cds. | 3712 | 137.25 |
| DQ014510.1 | contig_19509 | Eucalyptus grandis cellulose synthase 6 (CesA6) mRNA, complete cds. | 3782 | 97.32 |
| DQ227992.1 | contig_6857 | Eucalyptus grandis thioredoxin h mRNA, complete cds. | 354 | 133.93 |

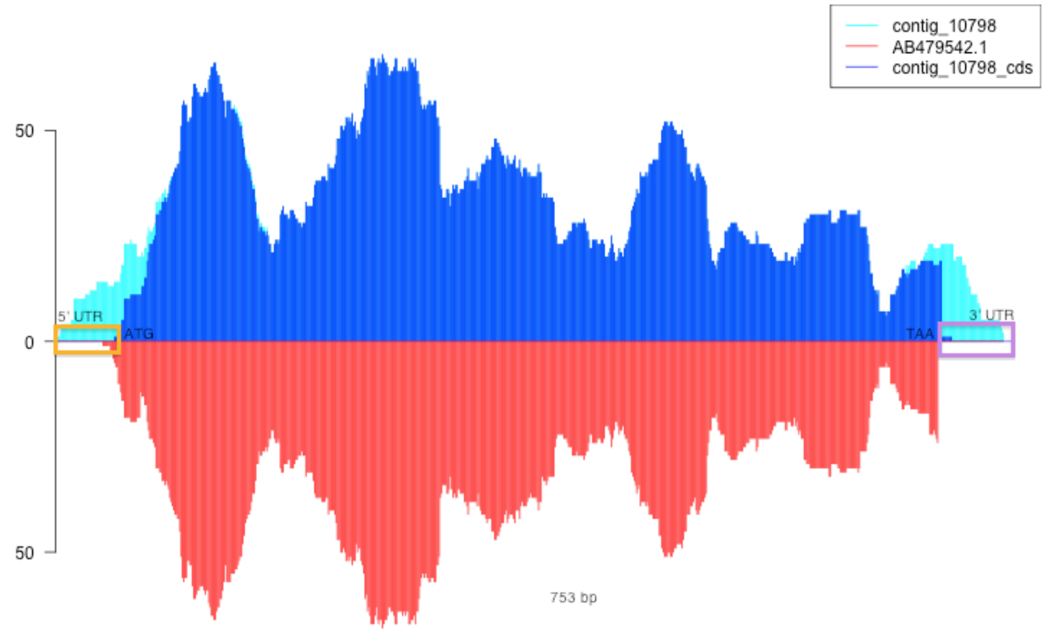
| | | | | |
|---|---|---|---|---|
| DQ227993.1 | contig_69050 | Eucalyptus grandis sucrose synthase (SuSy1) mRNA, complete cds. | 2498 | 250.38 |
| DQ227994.1 | contig_40644 | Eucalyptus grandis sucrose synthase (SuSy3) mRNA, complete cds. | 2508 | 220.28 |
| EF179384.1 | contig_24067 | Eucalyptus grandis UDP-glucose dehydrogenase (UGDH) mRNA, complete cds. | 1443 | 812.03 |
| EF534216.1 | contig_319 | Eucalyptus grandis fasciclin-like arabinogalactan protein (FLA1) mRNA, complete cds. | 1179 | 666.30 |
| EF534217.1 | contig_4434 | Eucalyptus grandis fasciclin-like arabinogalactan protein (FLA2) mRNA, complete cds. | 1125 | 180.66 |
| EF534218.1 | contig_2707 | Eucalyptus grandis fasciclin-like arabinogalactan protein (FLA3) mRNA, complete cds. | 1033 | 224.10 |
| EF534219.1 | contig_2477 | Eucalyptus grandis beta-tubulin (TUB1) mRNA, complete cds. | 1583 | 285.33 |
| EF534220.1 | contig_64905 | Eucalyptus grandis beta-tubulin (TUB2) mRNA, complete cds. | 1654 | 55.93 |
| EF534223.1 | contig_4441 | Eucalyptus grandis beta-tubulin (TUB5) mRNA, complete cds. | 1607 | 307.08 |
| EF534224.1 | contig_100 | Eucalyptus grandis alpha-tubulin (TUA1) mRNA, complete cds. | 1657 | 674.32 |
| EU737107.1 | contig_2692 | Eucalyptus grandis UTP-glucose 1 phosphate uridylyltransferase (UGP) mRNA, complete cds. | 1431 | 153.30 |
| EU737108.1 | contig_33128 | Eucalyptus grandis UDP-D-glucuronate carboxy-lyase (UXS1) mRNA, complete cds. | 1041 | 158.60 |
| EU770570.1 | contig_2246 | Eucalyptus grandis iron-sulfer cluster scaffold protein ISU1 (ISU1) mRNA, complete cds. | 756 | 78.07 |

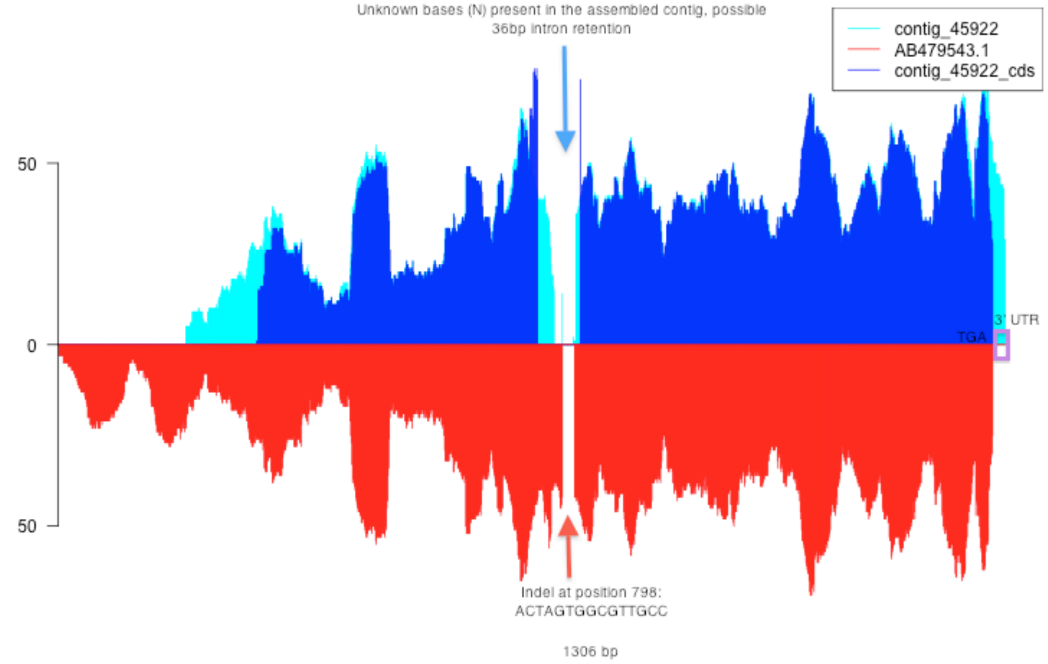
| | | | | |
|---|---|---|---|---|
| EU770571.1 | contig_31483 | Eucalyptus grandis iron-sulfer cluster scaffold protein NFU4 (NFU4) mRNA, partial cds. | 869 | 13.30 |
| EU770572.1 | contig_15010 | Eucalyptus grandis iron-sulfer cluster scaffold protein ISA1 (ISA1) mRNA, partial cds. | 822 | 25.81 |
| EU770573.1 | contig_25291 | Eucalyptus grandis iron-sulfer cluster scaffold protein NFS1 (NFS1) mRNA, partial cds. | 871 | 16.29 |

### C.1.2. Alignment coverage graphs of the 33 full length cDNA sequences and assembled contigs

Comparison of 33 *de novo*-assembled contigs of the *Eucalyptus grandis* x *Eucalyptus urophylla* clone compared to the reference contigs obtained from Sanger sequencing. Peak heights indicates the actual coverage per base (CPB) across the contig. The CBP of the assembled contig is shown in cyan, the CBP of the predicted CDS in dark blue, and the CPB of the reference sequence in red. Where present, the 5' UTR (orange box) and the 3' UTR (purple box) is indicated. Large gaps in the global alignment between the sequences are indicated by gaps in the graph, and possible reasons for the gap annotated on each graph. The graphs are also available as supplementary material for the article by Mizrachi *et al.* (2010).

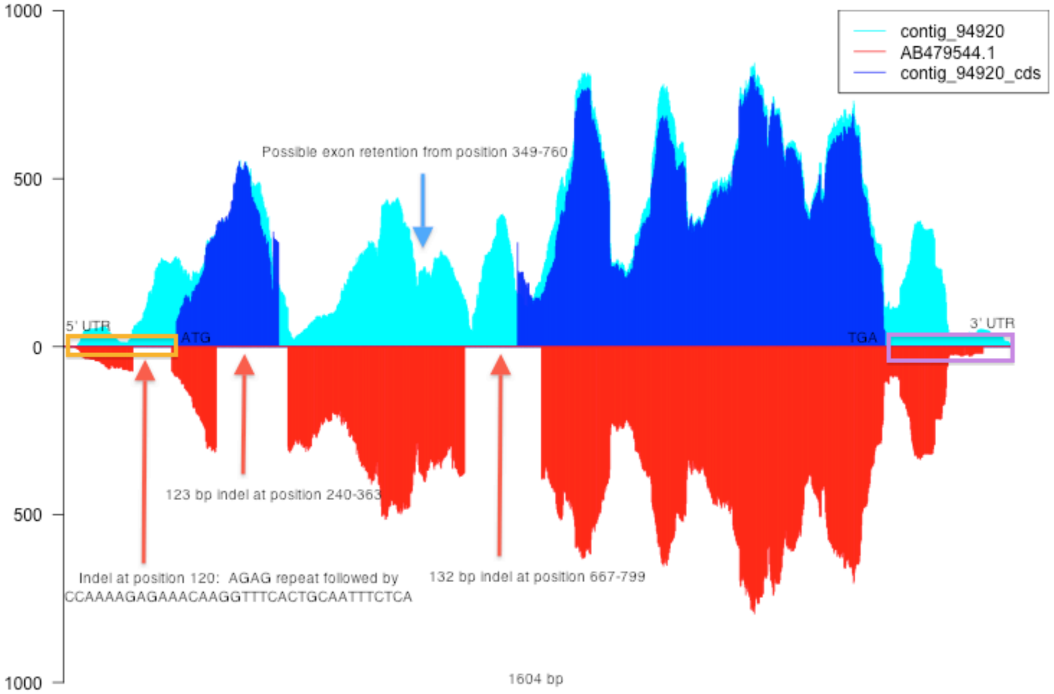

100
50
0
50
100
contig_87094
AB465730.1
contig_87094_cds
Indel at position 534: GTTGGA
5` UTR
ATG
TAG
3' UTR
985 bp


50
0
50
contig_10798
AB479542.1
contig_10798_cds
5' UTR
ATG
TAA
3' UTR
753 bp



Unknown bases (N) present in the assembled contig, possible 36bp intron retention

Indel at position 798: ACTAGTGGCGTTGCC

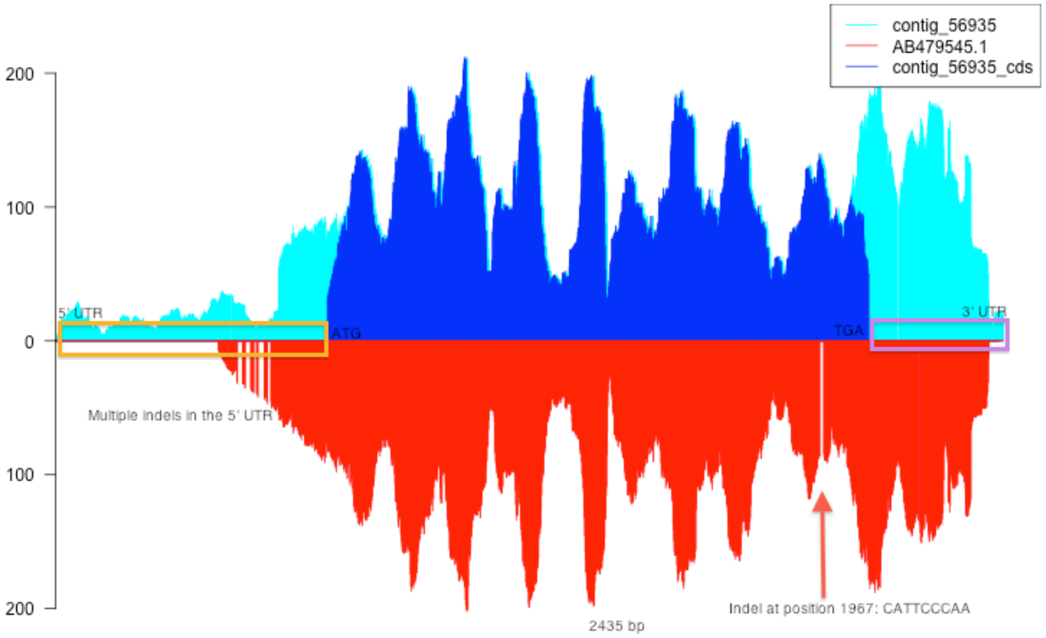


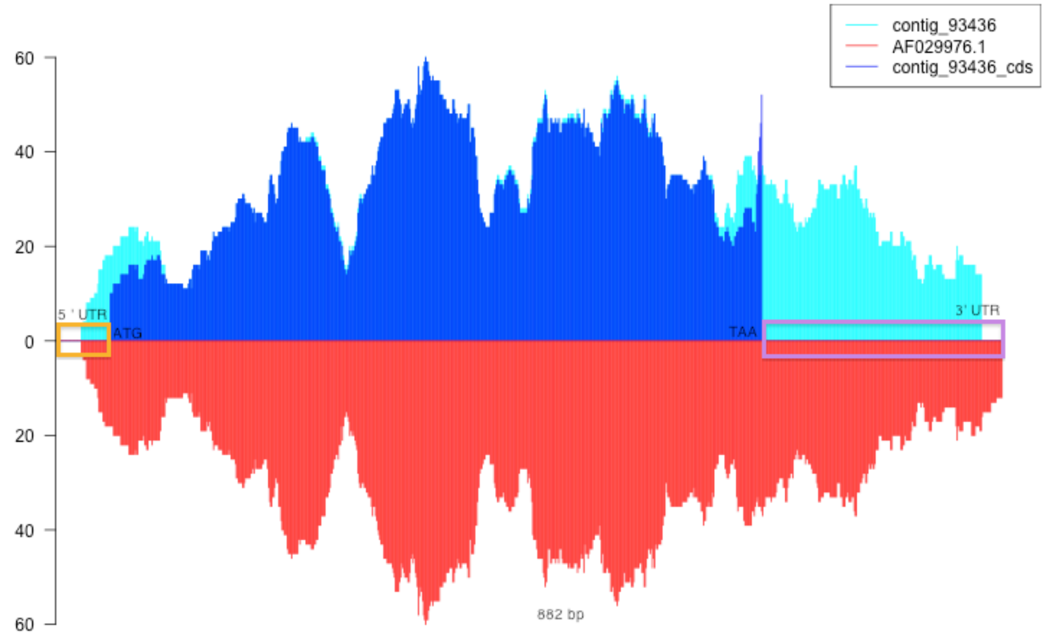
Possible exon retention from position 349-760

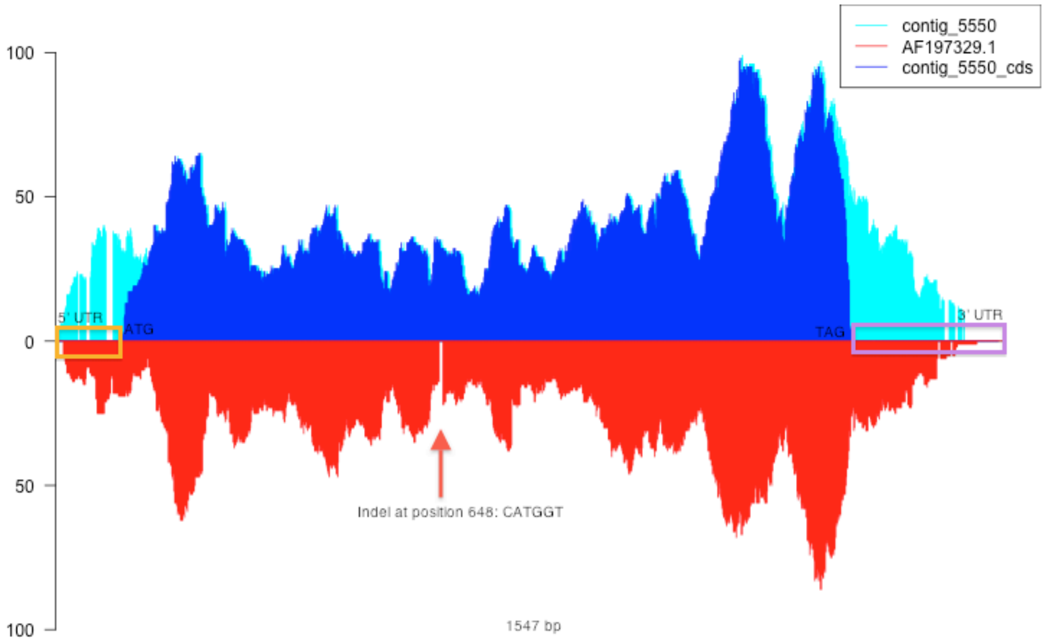
123 bp indel at position 240-363

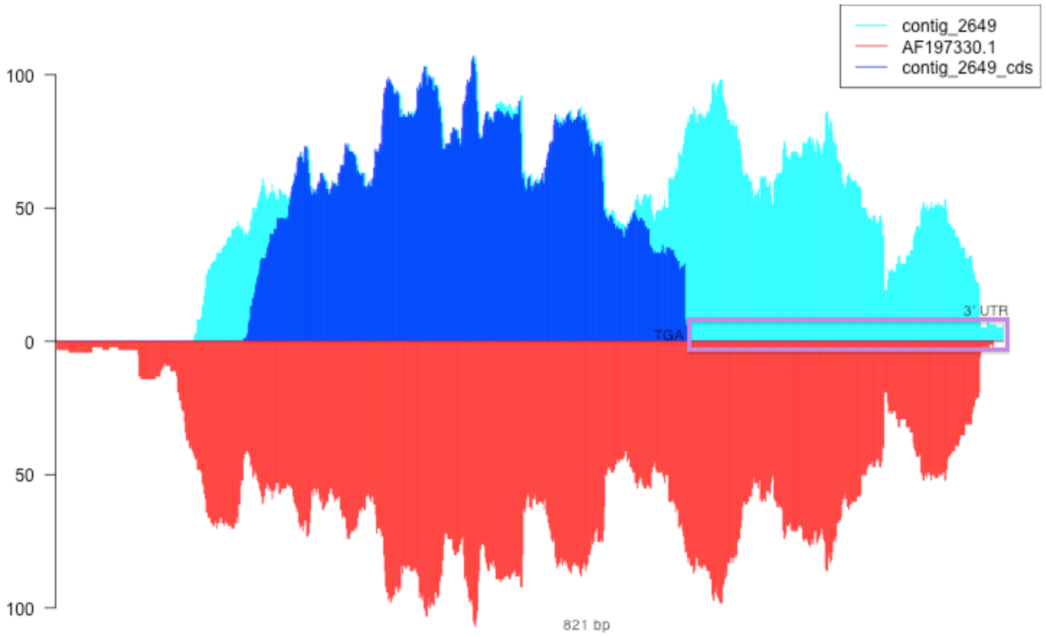
Indel at position 120: AGAG repeat followed by CCAAAAGAGAAACAAGGTTTCACTGCAATTTCTCA

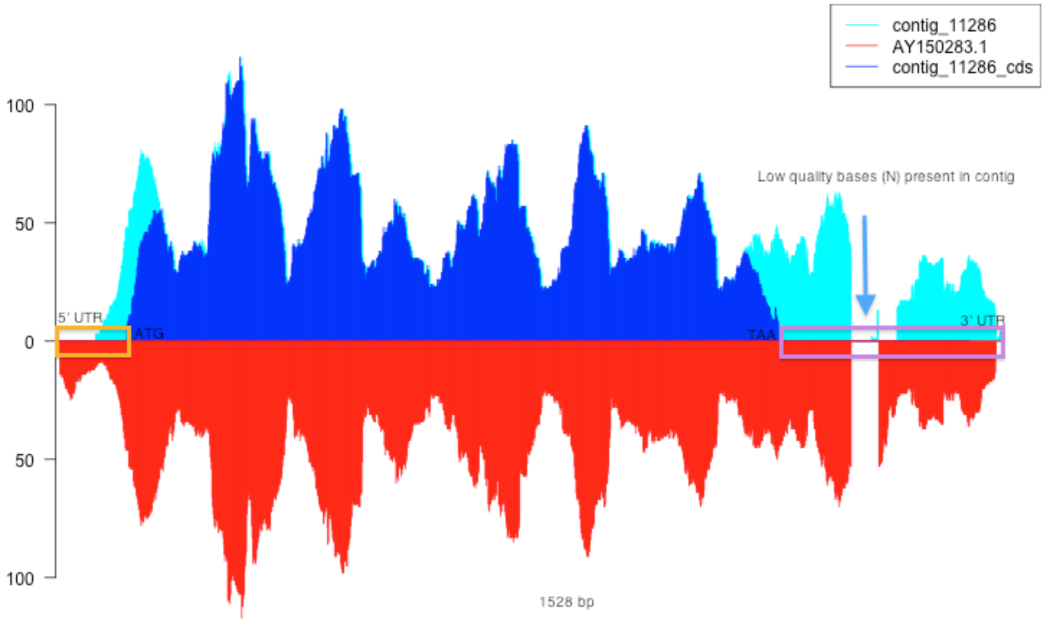
132 bp indel at position 667-799

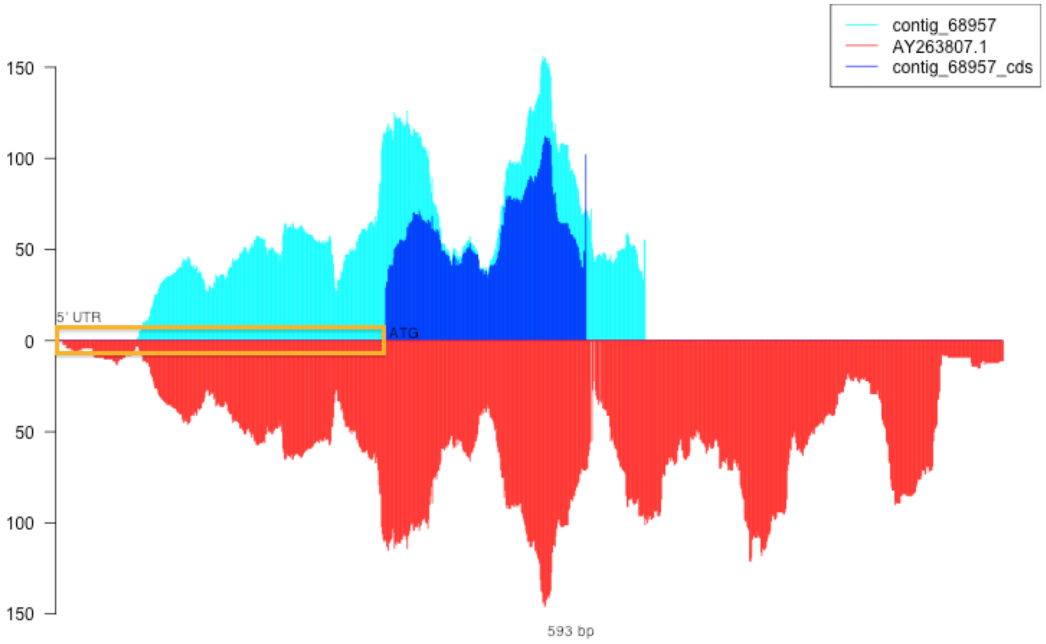

contig_56935
AB479545.1
contig_56935_cds
200
100
0
100
200
5' UTR
ATG
TGA
3' UTR
Multiple indels in the 5' UTR
Indel at position 1967: CATTCCCAA
2435 bp

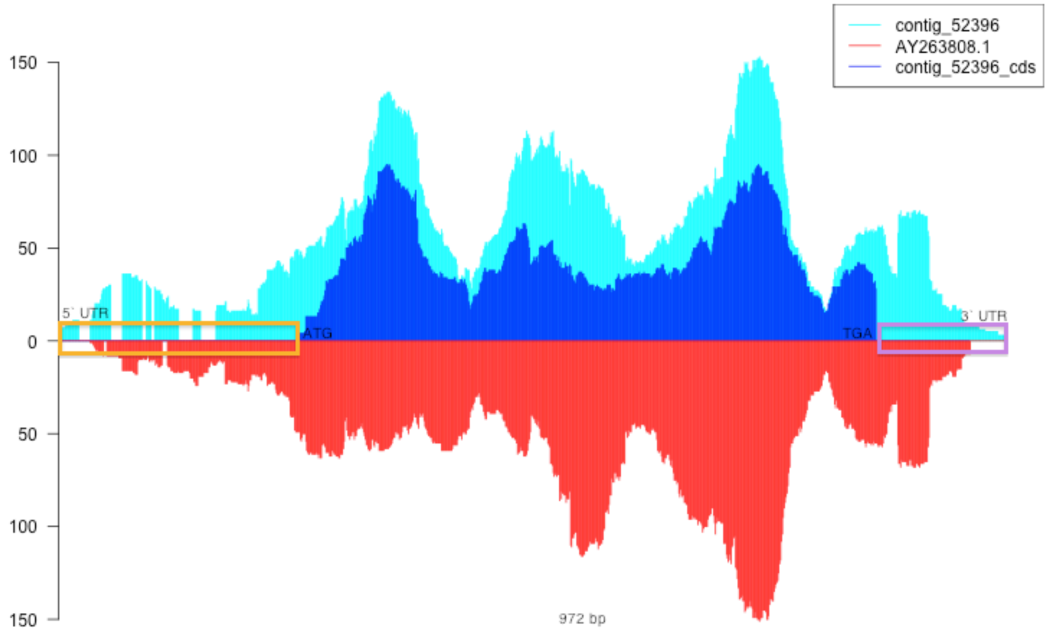

contig_93436
AF029976.1
contig_93436_cds
60
40
20
0
20
40
60
5 ' UTR
ATG
TAA
3' UTR
882 bp

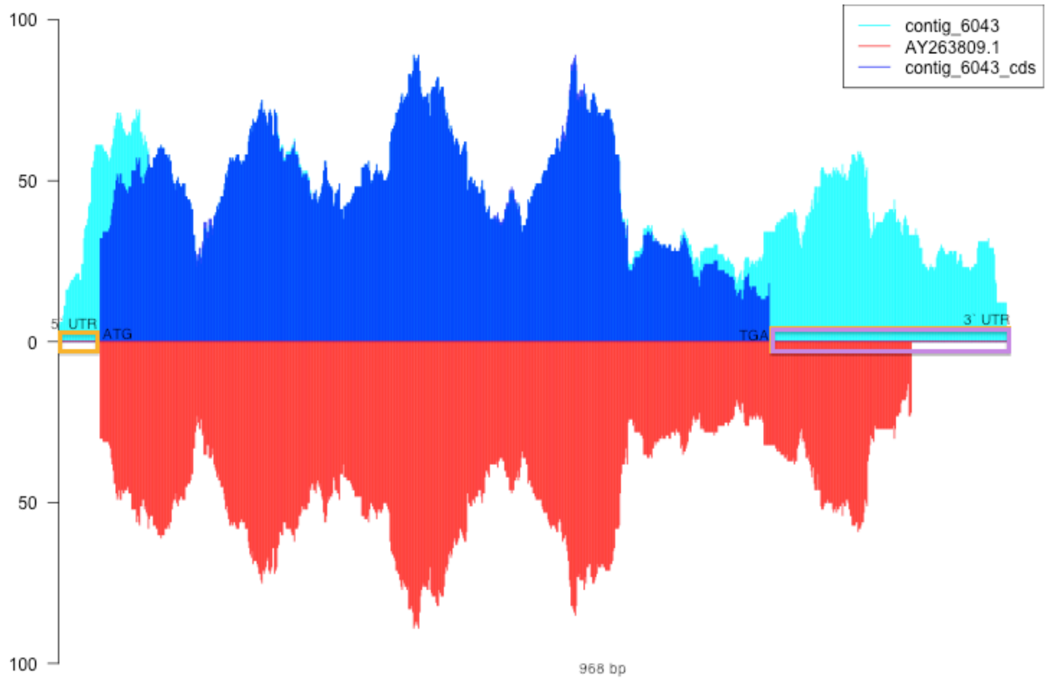

contig_5550
AF197329.1
contig_5550_cds
100
50
0
50
100
5' UTR
ATG
TAG
3' UTR
Indel at position 648: CATGGT
1547 bp

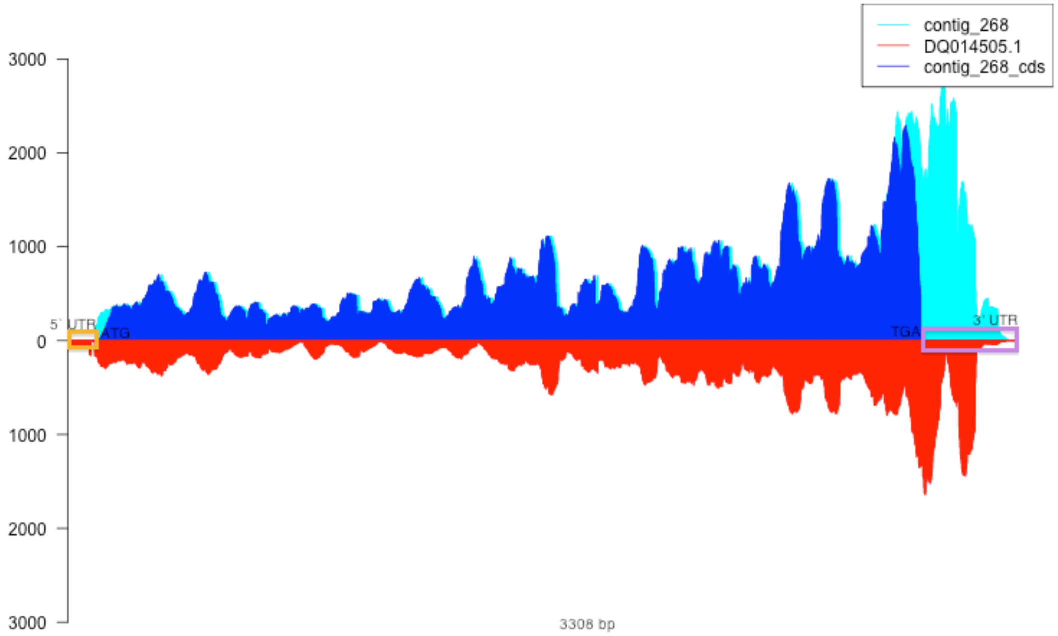

contig_2649
AF197330.1
contig_2649_cds
100
50
0
50
100
TGA
3' UTR
821 bp

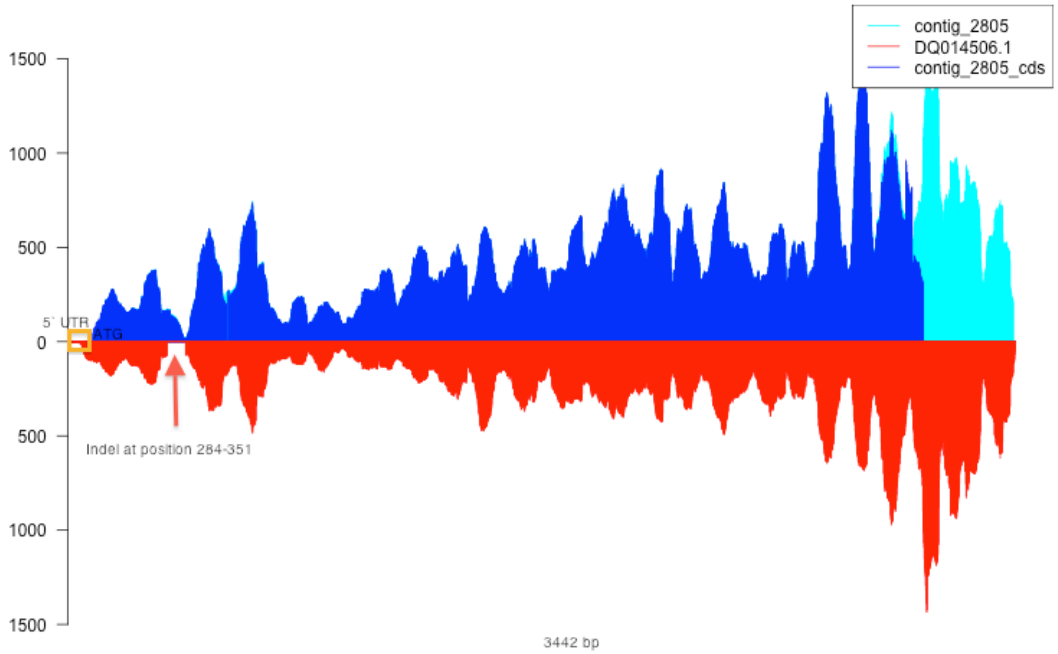

contig_11286
AY150283.1
contig_11286_cds
100
50
0
50
100
5' UTR
ATG
TAA
3' UTR
Low quality bases (N) present in contig
1528 bp

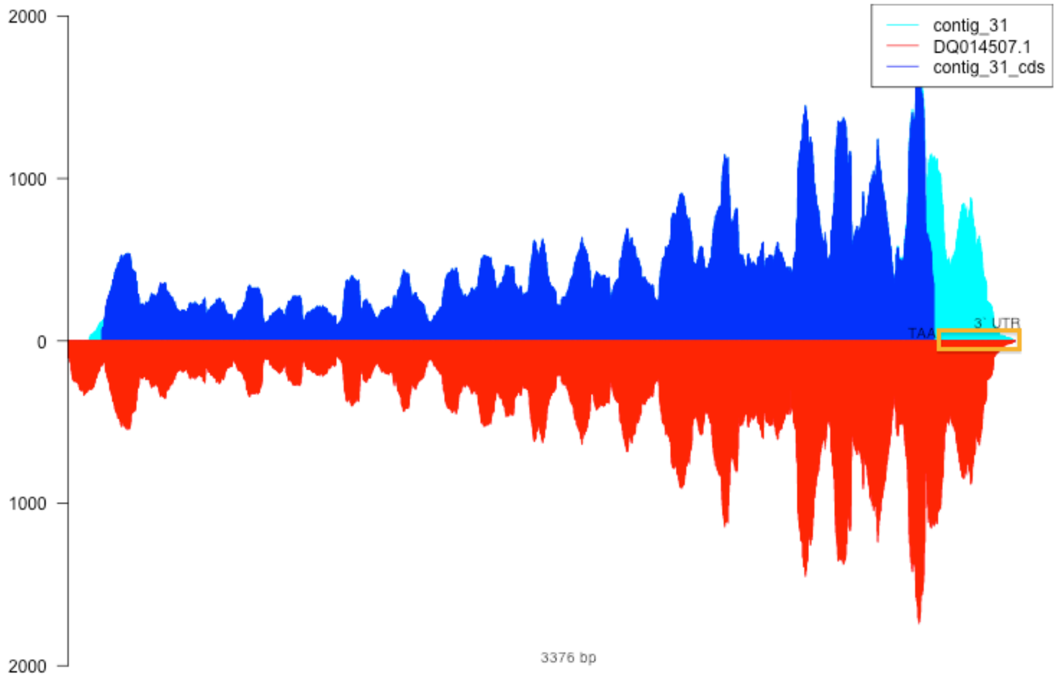

contig_68957
AY263807.1
contig_68957_cds
150
100
50
0
50
100
150
5' UTR
ATG
593 bp

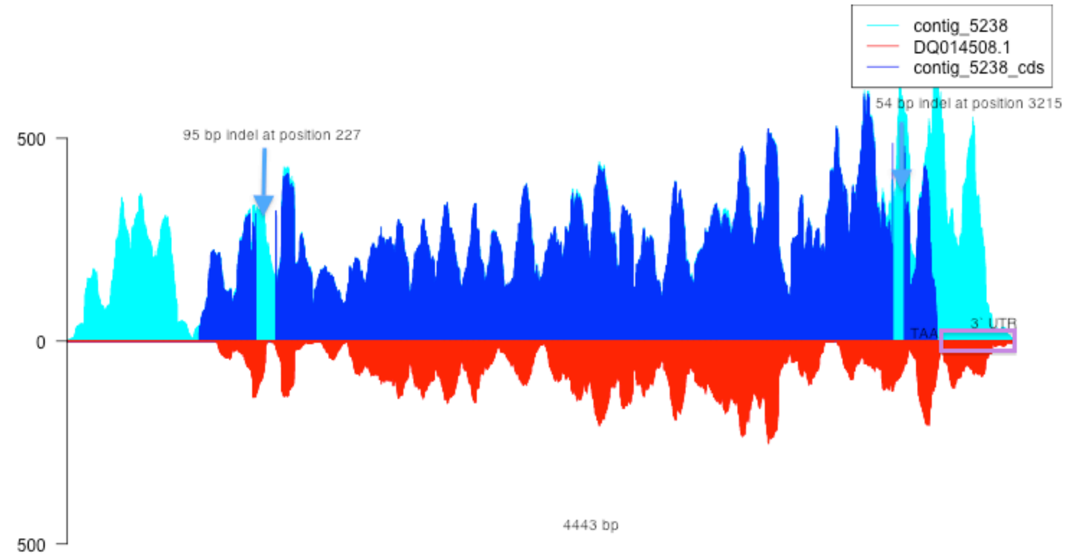

contig_52396
AY263808.1
contig_52396_cds
150
100
50
0
50
100
150
5` UTR
ATG
TGA
3` UTR
972 bp

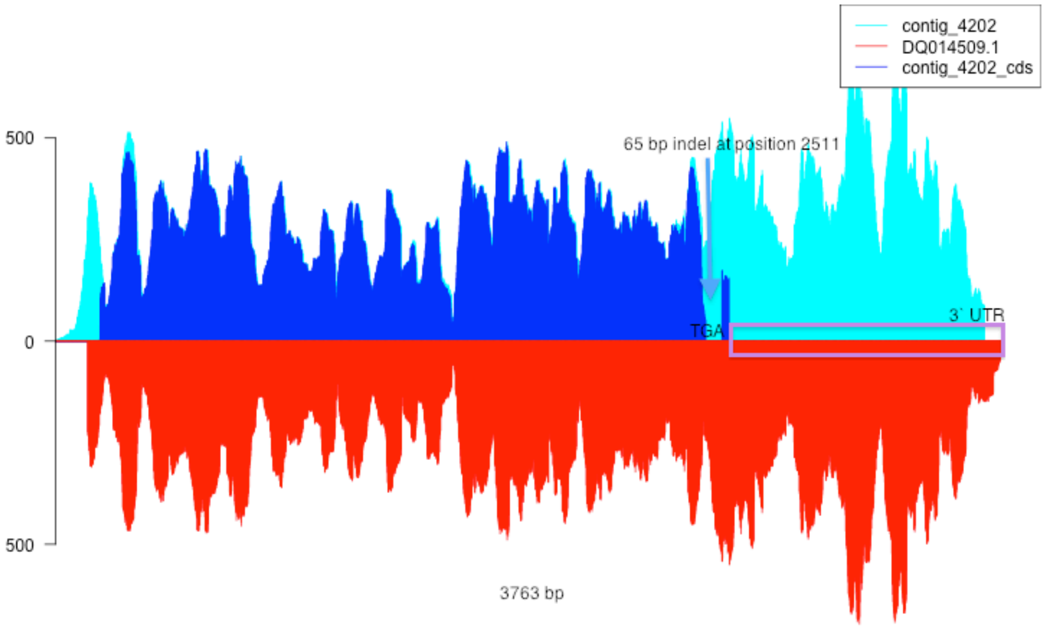

contig_6043
AY263809.1
contig_6043_cds
100
50
0
50
100
5` UTR
ATG
TGA
3` UTR
968 bp

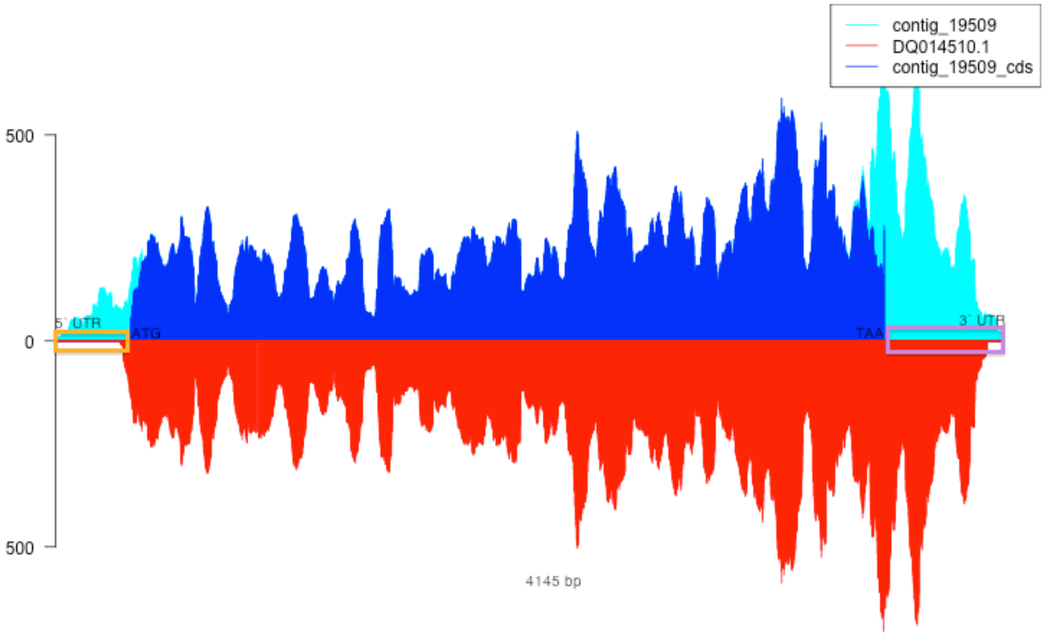

contig_268
DQ014505.1
contig_268_cds
3000
2000
1000
0
1000
2000
3000
5` UTR
ATG
TGA
3' UTR
3308 bp

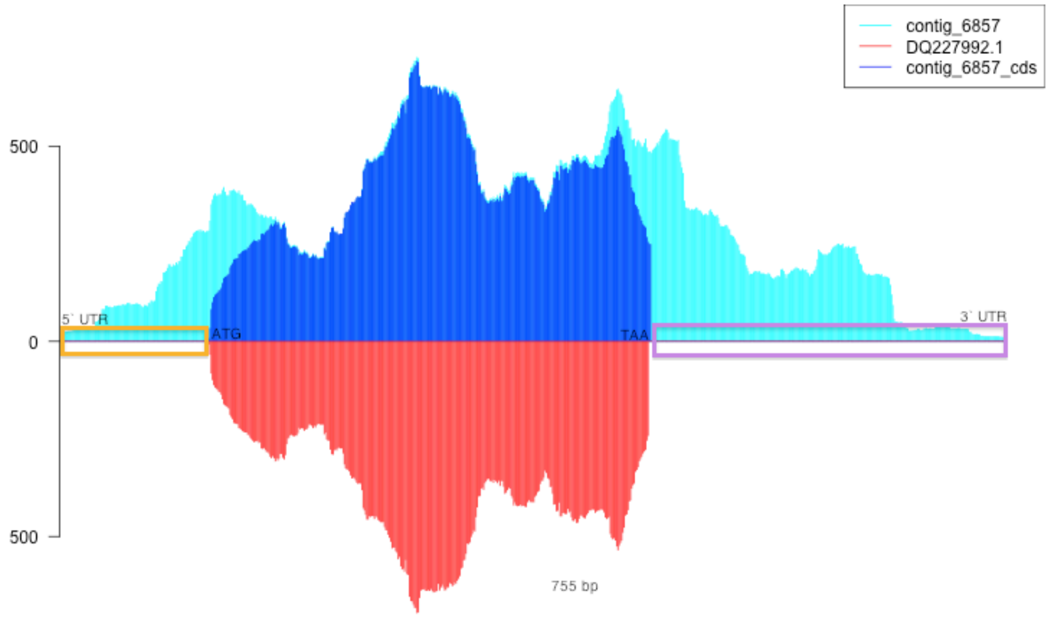

contig_2805
DQ014506.1
contig_2805_cds
1500
1000
500
0
500
1000
1500
5` UTR
ATG
Indel at position 284-351
3442 bp

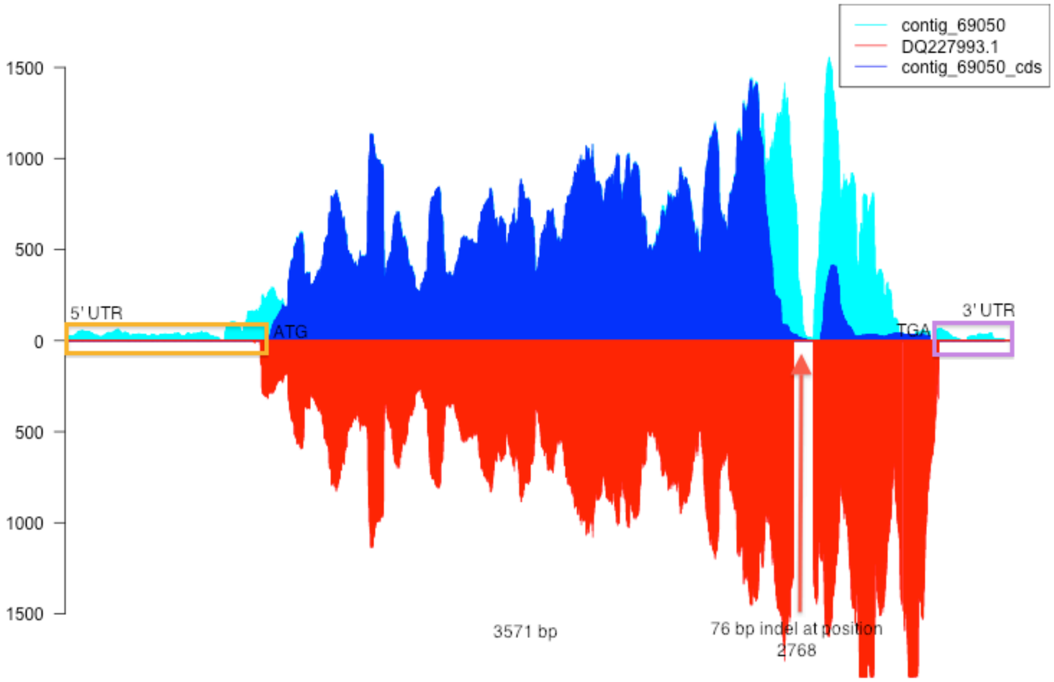

2000
1000
0
1000
2000
contig_31
DQ014507.1
contig_31_cds
TAA
3' UTR
3376 bp

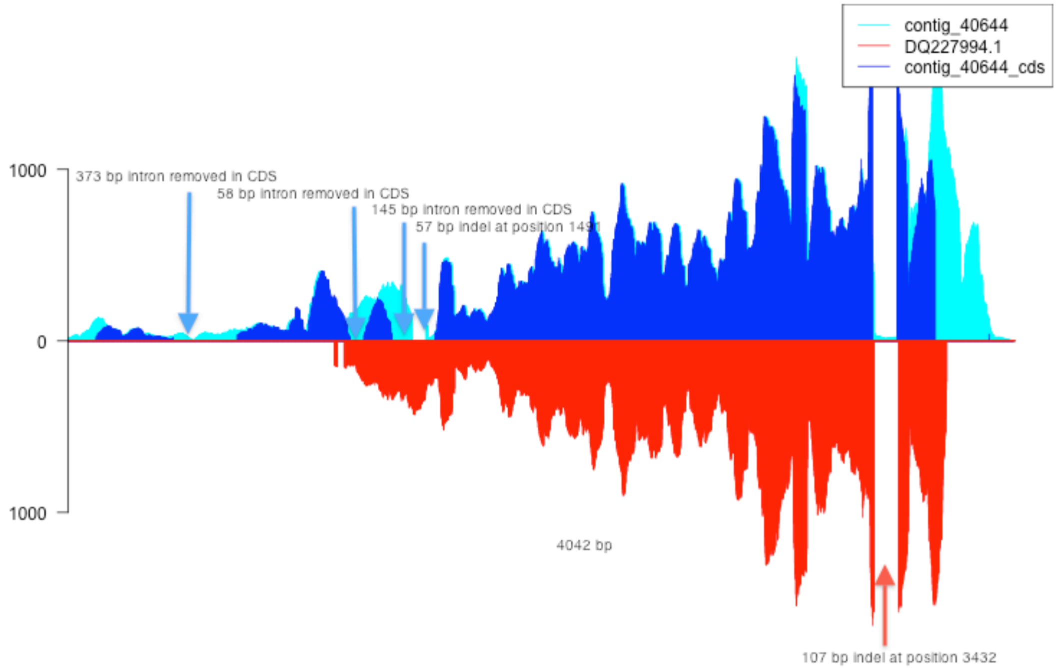

500
0
500
contig_5238
DQ014508.1
contig_5238_cds
95 bp indel at position 227
54 bp indel at position 3215
TAA
3' UTR
4443 bp

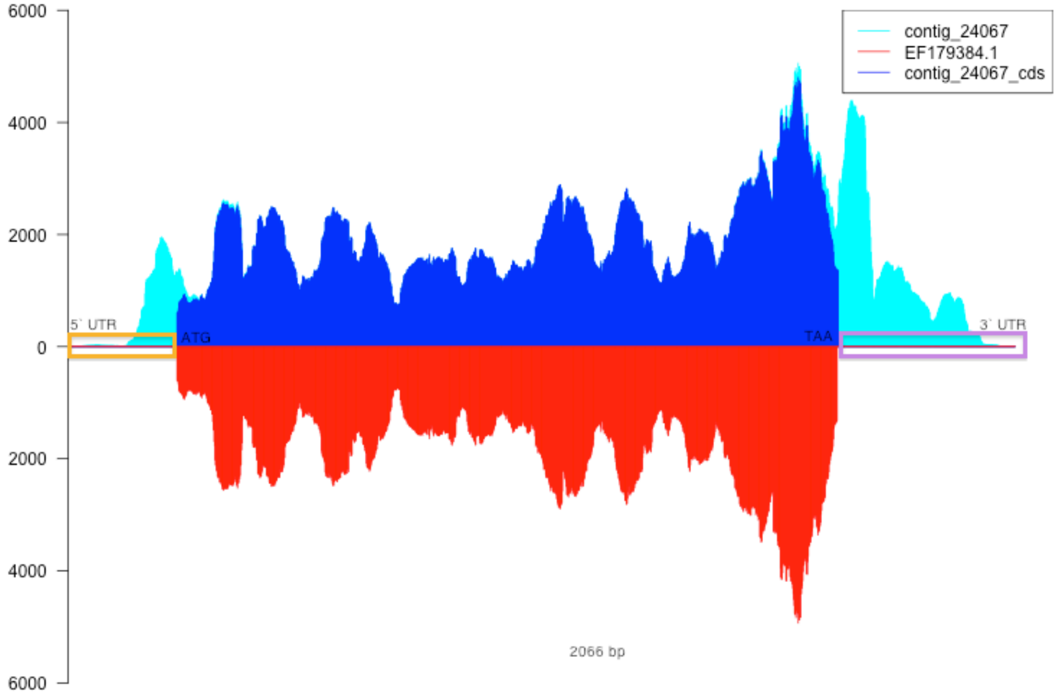

contig_4202
DQ014509.1
contig_4202_cds
500
0
500
65 bp indel at position 2511
TGA
3` UTR
3763 bp

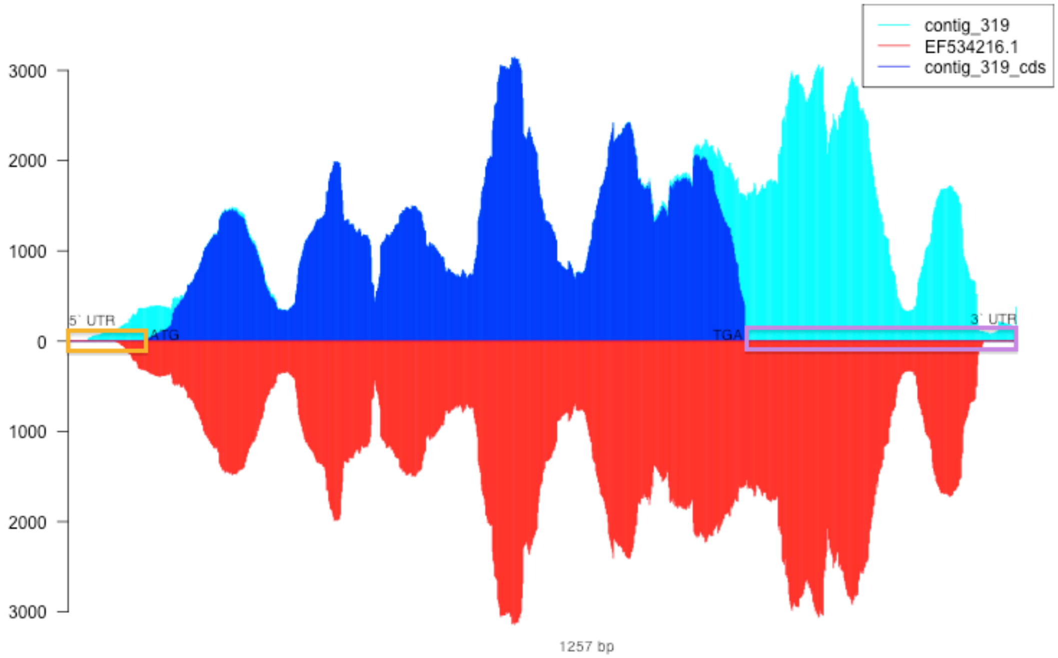

contig_19509
DQ014510.1
contig_19509_cds
500
0
500
5` UTR
ATG
TAA
3` UTR
4145 bp

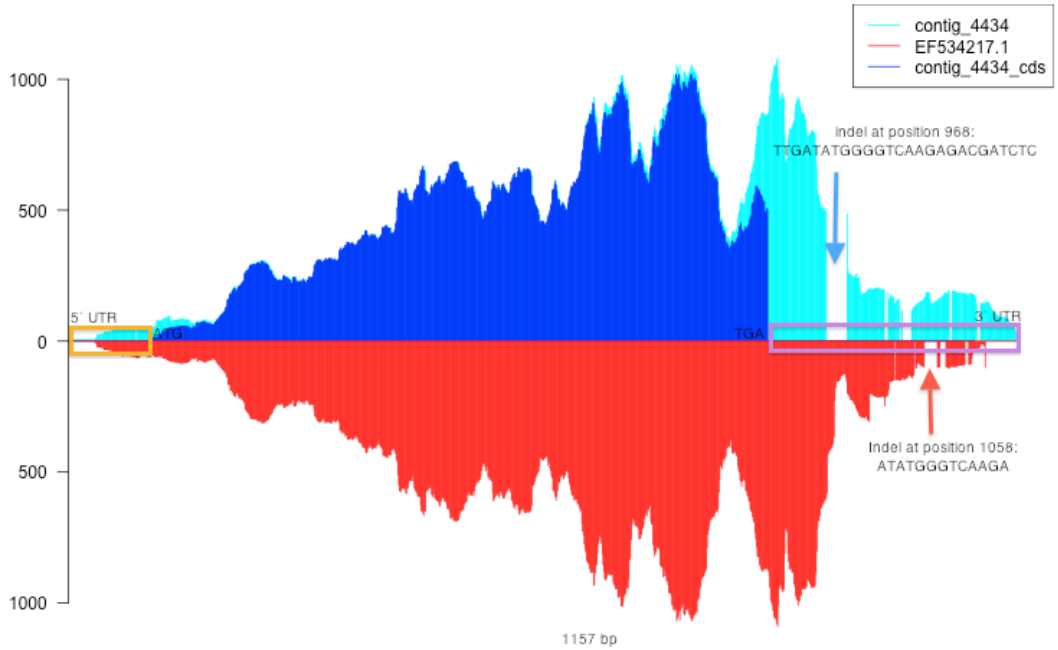

contig_6857
DQ227992.1
contig_6857_cds
500
0
500
5` UTR
ATG
TAA
3` UTR
755 bp

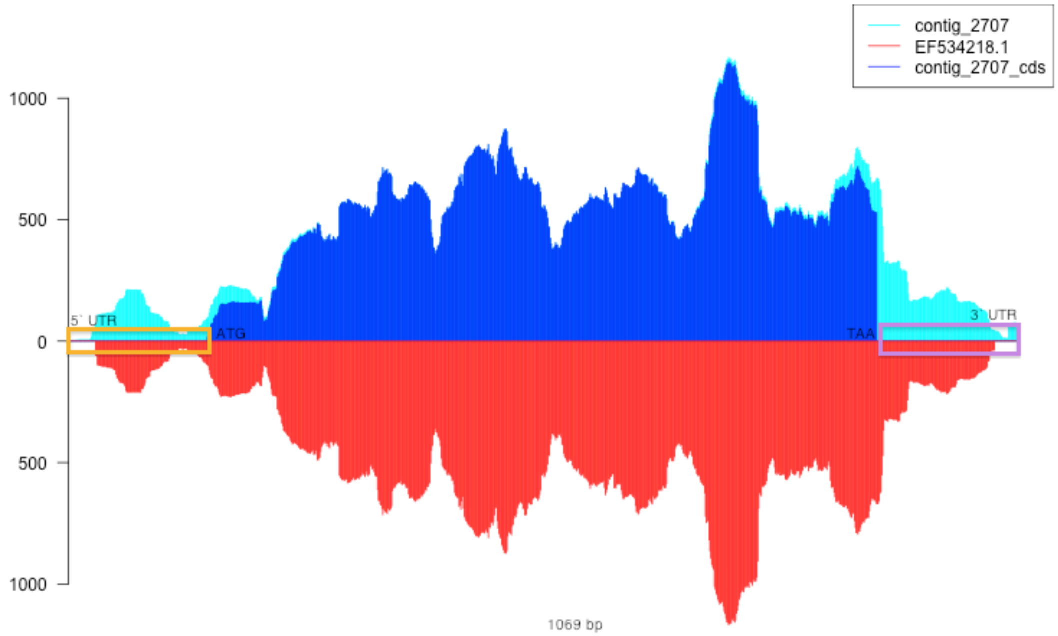

contig_69050
DQ227993.1
contig_69050_cds
1500
1000
500
0
500
1000
1500
5' UTR
ATG
TGA
3' UTR
3571 bp
76 bp indel at position 2768

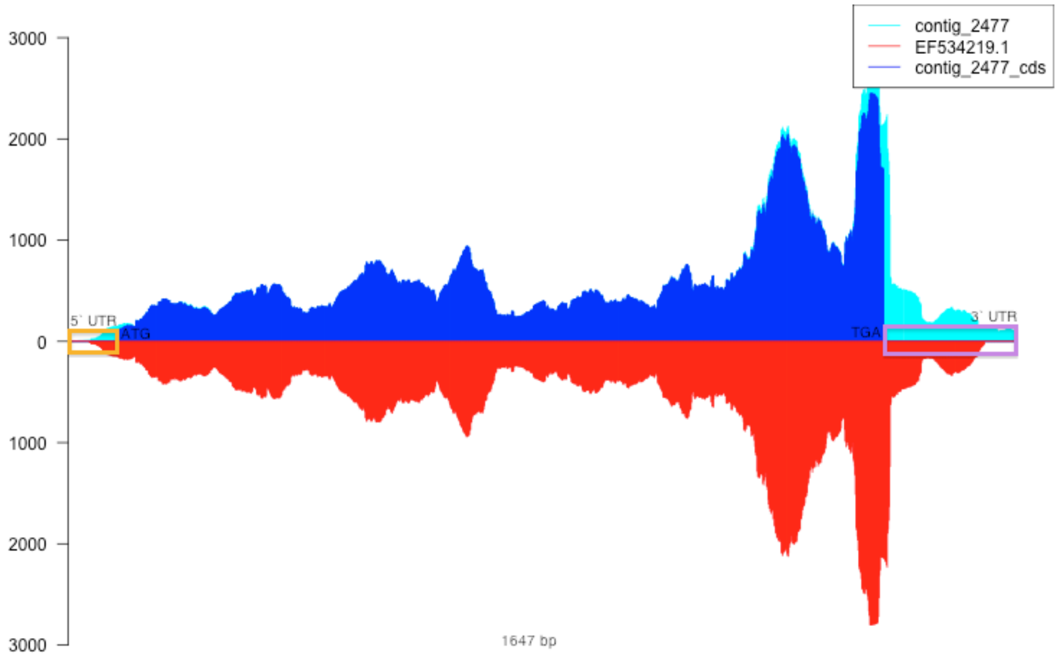

contig_40644
DQ227994.1
contig_40644_cds
1000
0
1000
373 bp intron removed in CDS
58 bp intron removed in CDS
145 bp intron removed in CDS
57 bp indel at position 1491
4042 bp
107 bp indel at position 3432

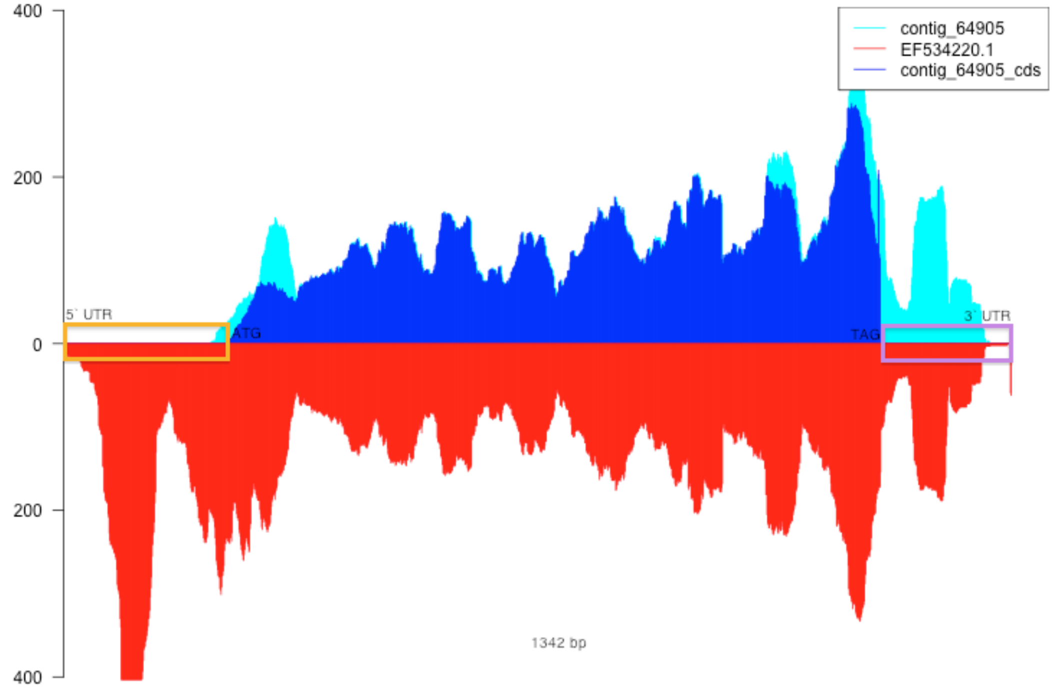

contig_24067
EF179384.1
contig_24067_cds
6000
4000
2000
0
2000
4000
6000
5' UTR
ATG
TAA
3' UTR
2066 bp

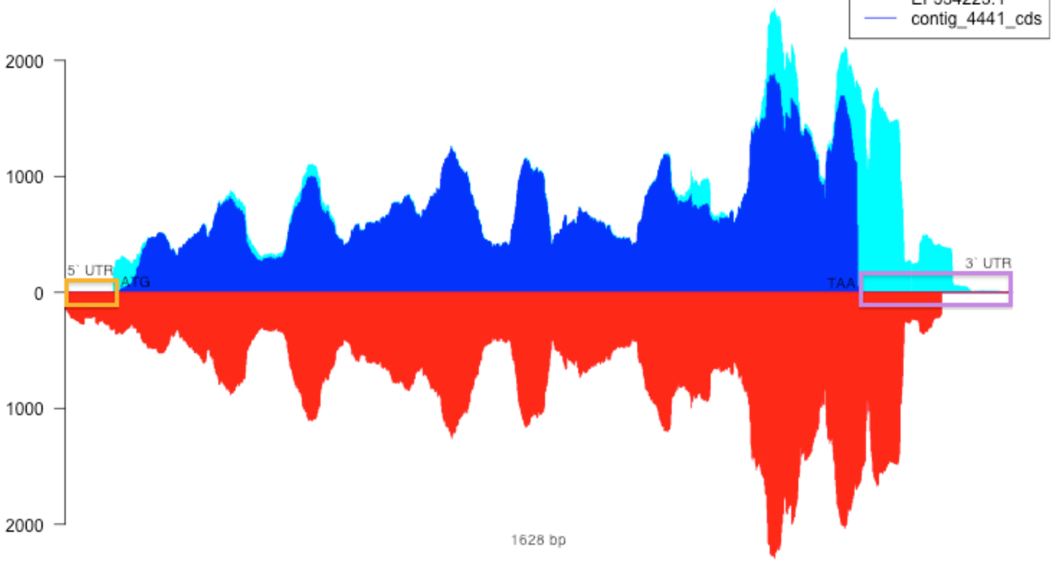

contig_319
EF534216.1
contig_319_cds
3000
2000
1000
0
1000
2000
3000
5` UTR
ATG
TGA
3` UTR
1257 bp

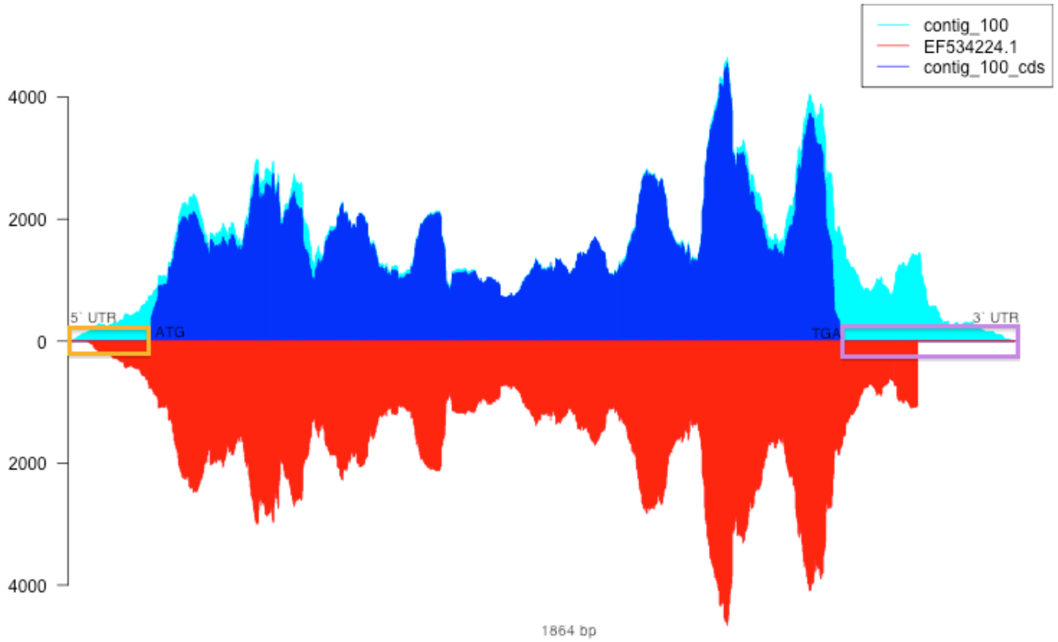

contig_4434
EF534217.1
contig_4434_cds
1000
500
0
500
1000
5` UTR
ATG
TGA
3` UTR
Indel at position 968:
TTGATATGGGGTCAAGAGACGATCTC
Indel at position 1058:
ATATGGGTCAAGA
1157 bp

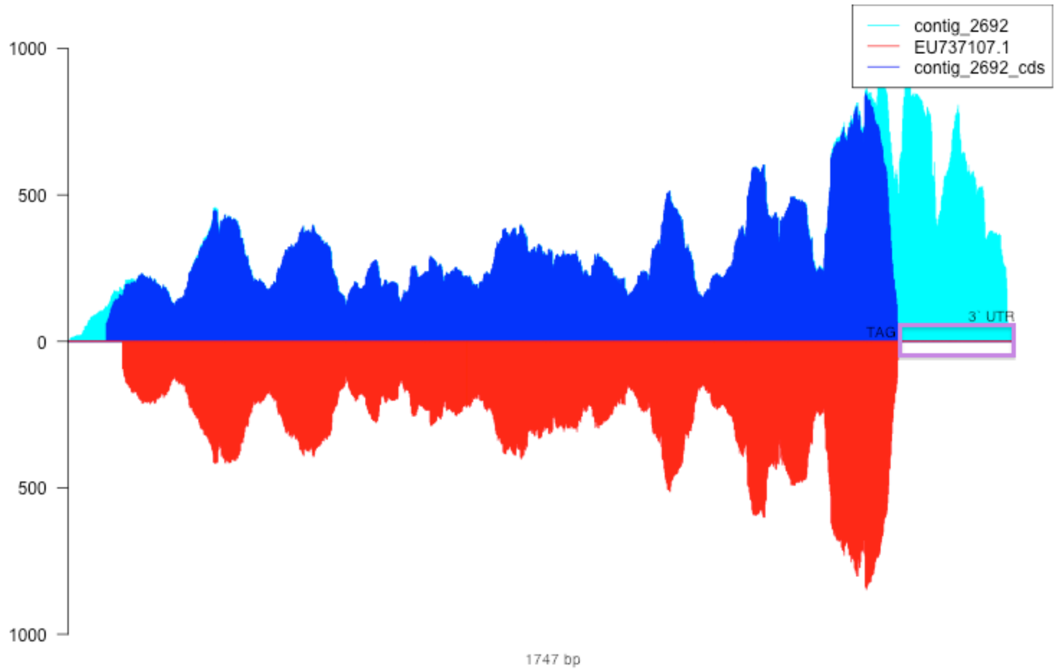

contig_2707
EF534218.1
contig_2707_cds
1000
500
0
500
1000
5` UTR
ATG
TAA
3` UTR
1069 bp

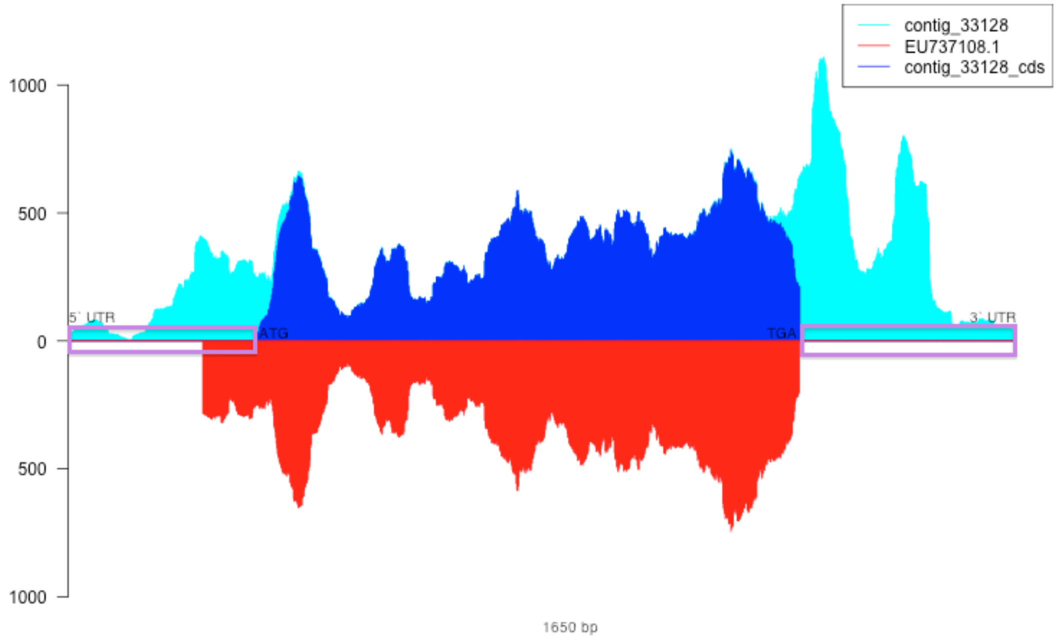

contig_2477
EF534219.1
contig_2477_cds
3000
2000
1000
0
1000
2000
3000
5` UTR
ATG
TGA
3` UTR
1647 bp

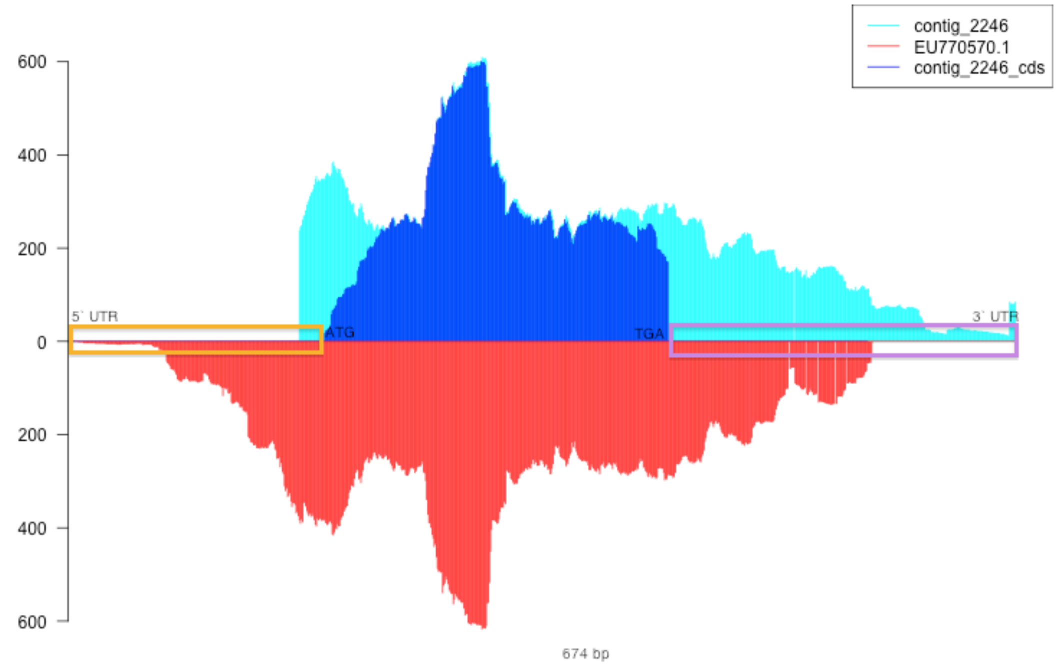

400
200
0
200
400
contig_64905
EF534220.1
contig_64905_cds
5` UTR
ATG
TAG
3` UTR
1342 bp

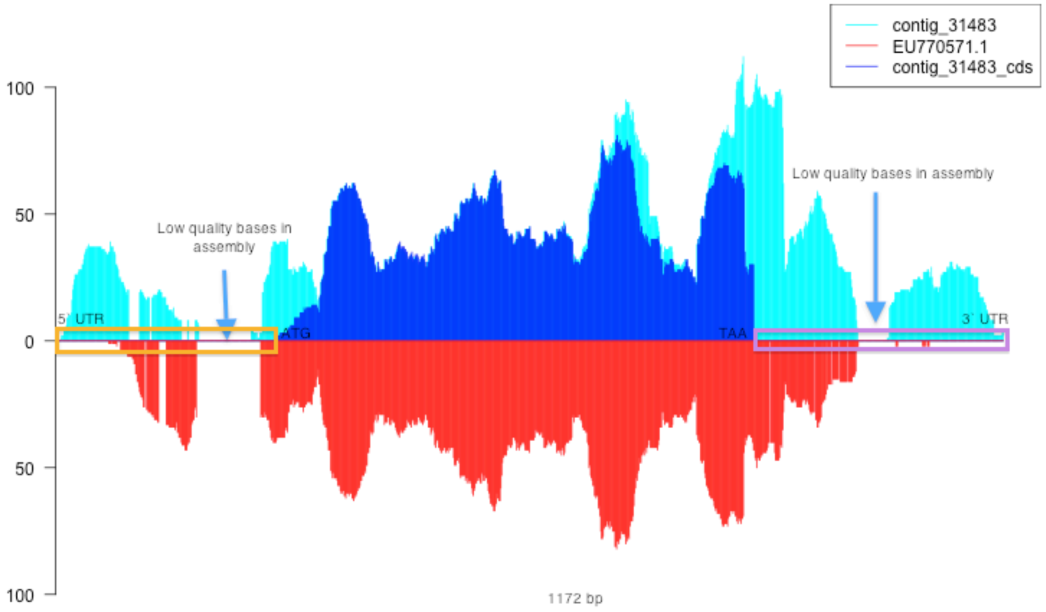

2000
1000
0
1000
2000
contig_4441
EF534223.1
contig_4441_cds
5` UTR
ATG
TAA
3` UTR
1628 bp

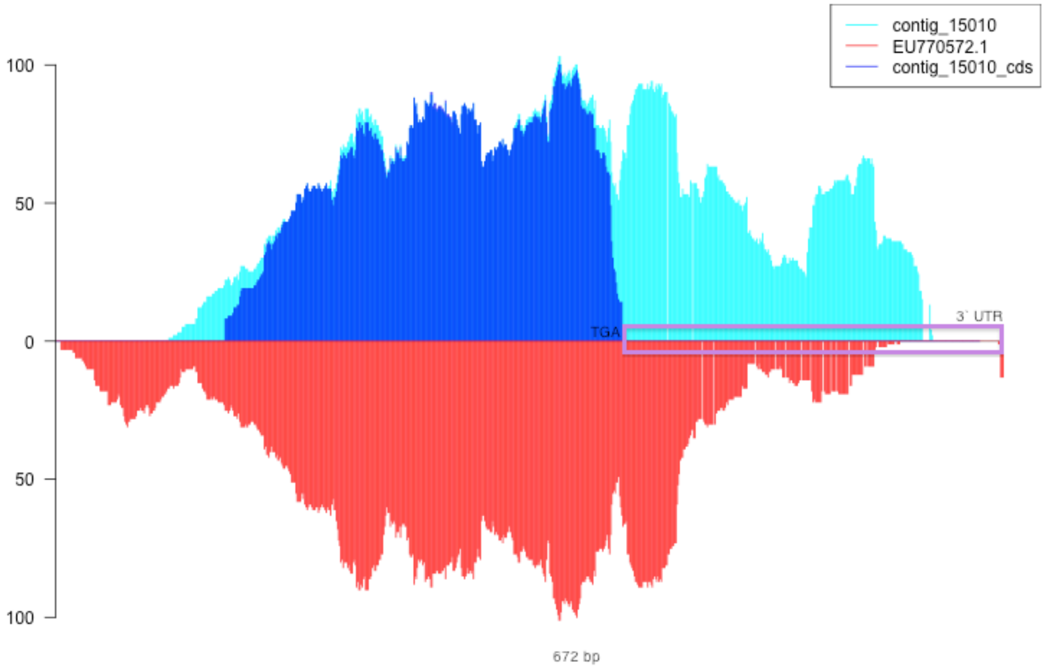

contig_100
EF534224.1
contig_100_cds
4000
2000
0
2000
4000
5` UTR
ATG
TGA
3` UTR
1864 bp

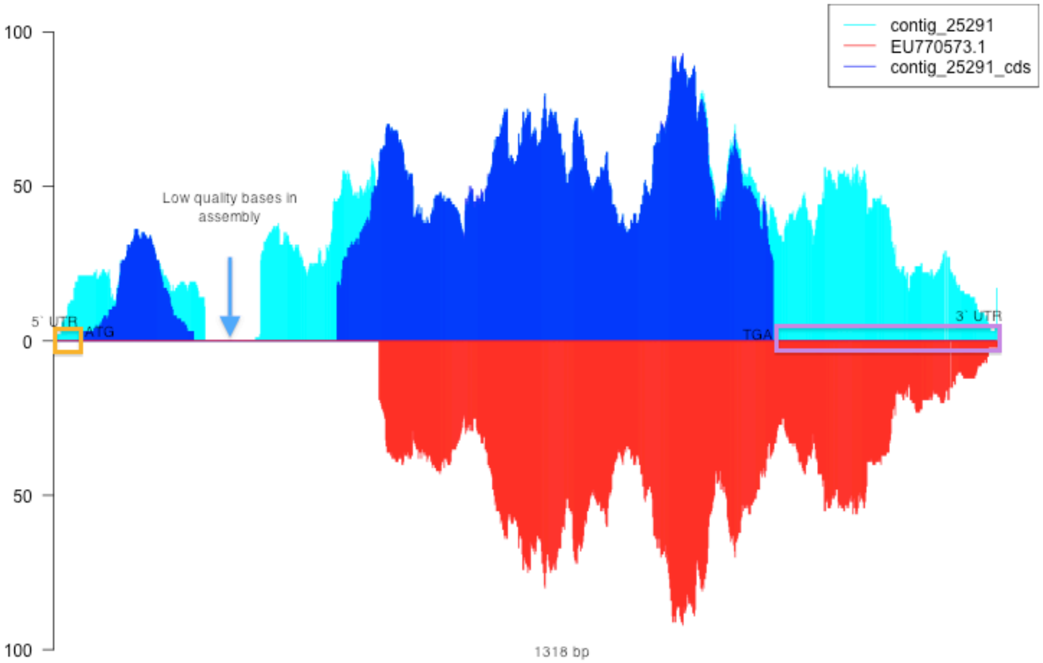

contig_2692
EU737107.1
contig_2692_cds
1000
500
0
500
1000
TAG
3` UTR
1747 bp


contig_33128
EU737108.1
contig_33128_cds
1000
500
0
500
1000
5` UTR
ATG
TGA
3` UTR
1650 bp


contig_2246
EU770570.1
contig_2246_cds
600
400
200
0
200
400
600
5` UTR
ATG
TGA
3` UTR
674 bp


contig_31483
EU770571.1
contig_31483_cds
100
50
0
50
100
Low quality bases in assembly
Low quality bases in assembly
5` UTR
ATG
TAA
3` UTR
1172 bp


contig_15010
EU770572.1
contig_15010_cds
100
50
0
50
100
TGA
3` UTR
672 bp


contig_25291
EU770573.1
contig_25291_cds
100
50
0
50
100
Low quality bases in assembly
5` UTR
ATG
TGA
3` UTR
1318 bp

### C.1.3. Alignment of contig 68291 before and after extension

The complete alignment of contig or node 68291 before and after the coverage-assisted re-assembly of the dataset. The aligment was performed with the the `ClustalW` program, and no editing of the alignment was performed. The alignment shows that although 1 485 bases was reportedly added to the contig during extension, these bases mostly consisted of the extension of a low quality region containing Ns. The extension did however resolve a 88 bp region of these low quality bases. The contig after extension also showed removed regions at the start and end of the original contig, due to the presence of a polyA region at the beginning of the sequence. An extract from the alignment is presented in Figure 3.6.

```
NODE_68291_before TTTTTTTTTTTTTTTTTTTTTTA[illegible]AGAGAGGCTTGGAGTGGTCAGGACTCTTA    80
NODE_68291_after  --------------------------------------------------------------------------------

                                                                              **    * * ***  * *
NODE_68291_before TTCCAGAGAGAAAGCATCAGCGCGGGCTGTCAAAGCTTCAATGAACGTATAGTTTGTTACTGAAGCGGAGATAAAGGAGA   160
NODE_68291_after  -----------------------------------------------------------AAAAAAAAAAAAAAAAAAAAA    21

                    **     ***********************************************************************
NODE_68291_before GTAAGNNNNCAAGCAAAGCTGGTGCTTTTTGTGCATCTCCATTGACTTGGCCATTTGGTCCATTGATAAGTTCGGCAACT   240
NODE_68291_after  AAAAAAAAACAAGCAAAGCTGGTGCTTTTTGTGCATCTCCATTGACTTGGCCATTTGGTCCATTGATAAGTTCGGCAACT   101

                  ********************************************************************************
NODE_68291_before CAGGCGATTCCTGGCTTTGCGTCGGTGAGTTCTTCGTGATCTGGCAATGGCGTCGGCTCTAGCCGGCGATGATTTGGCTA   320
NODE_68291_after  CAGGCGATTCCTGGCTTTGCGTCGGTGAGTTCTTCGTGATCTGGCAATGGCGTCGGCTCTAGCCGGCGATGATTTGGCTA   181

                  ********************************************************************************
NODE_68291_before GATCCACGAGCAGCCGCCGGAGCTGGGCCTCCGGGAGCCACCGGAGCTGGGCCTCCACGAGCTTCCGGGAGGCGTGGAAC   400
NODE_68291_after  GATCCACGAGCAGCCGCCGGAGCTGGGCCTCCGGGAGCCACCGGAGCTGGGCCTCCACGAGCTTCCGGGAGGCGTGGAAC   261

                  ********************************************************************************
NODE_68291_before GGCCCGCCAGATGTGTTCGCGCGGAGCGGGAGGCAGGACGACGAGGAGGAGCTCCGGTGGGCCGCCATCGAACGGCTGCC   480
NODE_68291_after  GGCCCGCCAGATGTGTTCGCGCGGAGCGGGAGGCAGGACGACGAGGAGGAGCTCCGGTGGGCCGCCATCGAACGGCTGCC   341

                  ********************************************************************************
NODE_68291_before AACGTATGACCGCCTCCGAAAAGGCATGCTGAAGCAAGTACTTGATACTGGGAGGGTGGTCCAGCAAGAAGTGGACGTGA   560
NODE_68291_after  AACGTATGACCGCCTCCGAAAAGGCATGCTGAAGCAAGTACTTGATACTGGGAGGGTGGTCCAGCAAGAAGTGGACGTGA   421

                  ********************************************************************************
NODE_68291_before CCAACCTCGGAATGCAGGACAAGAAGCAGTTGATGGAGAGCATCCTTAAGGTTGCGGAAGAAGACAATGAGAGGTTCTTG   640
NODE_68291_after  CCAACCTCGGAATGCAGGACAAGAAGCAGTTGATGGAGAGCATCCTTAAGGTTGCGGAAGAAGACAATGAGAGGTTCTTG   501

                  ********************************************************************************
NODE_68291_before AGGAGATTGAGAGACAGGACTGATAGGGTCGGGATCGAAATTCCGAAGATCGAAGTCCGGTGTGAGCATTTATCTGTAGA   720
NODE_68291_after  AGGAGATTGAGAGACAGGACTGATAGGGTCGGGATCGAAATTCCGAAGATCGAAGTCCGGTGTGAGCATTTATCTGTAGA   581

                  ********************************************************************************
NODE_68291_before AGGAGACGTGTACGTTGGAAGCAGAGCTCTCCCTACCCTTCTCAATGCCACTATGAACGCGATAGAGAGTGTTCTTGGAC   800
NODE_68291_after  AGGAGACGTGTACGTTGGAAGCAGAGCTCTCCCTACCCTTCTCAATGCCACTATGAACGCGATAGAGAGTGTTCTTGGAC   661

                  ********************************************************************************
NODE_68291_before TTATTCGGCTAGCCCCATCGAAGAAGAGAAAAATTCAGATACTTAAGGACGTGAACGGATTAGTCAGGCCTTCGAGGATG   880
NODE_68291_after  TTATTCGGCTAGCCCCATCGAAGAAGAGAAAAATTCAGATACTTAAGGACGTGAACGGATTAGTCAGGCCTTCGAGGATG   741

                  ********************************************************************************
NODE_68291_before ACCCTACTTTTGGGTCCACCGGGAGCTGGGAAGACAACATTGTTGCTGGCACTTGCTGGGAAACTAGACAGCGATCTGAG   960
NODE_68291_after  ACCCTACTTTTGGGTCCACCGGGAGCTGGGAAGACAACATTGTTGCTGGCACTTGCTGGGAAACTAGACAGCGATCTGAG   821

                  ********************************************************************************
NODE_68291_before GGTAACGGGAAAAGTCACCTACTGTGGTCACGAGCTAAACGAATTTGTTCCTCAAAGGACTTGCGCTTATATCAGCCAAC  1040
NODE_68291_after  GGTAACGGGAAAAGTCACCTACTGTGGTCACGAGCTAAACGAATTTGTTCCTCAAAGGACTTGCGCTTATATCAGCCAAC   901

                  ********************************************************************************
NODE_68291_before ATGATCTTCACTATGGGGAAATGACAGTTAGAGAGACATTGGACTTCTCGGGTCGCTGTTTGGGTGTAGGGACAAGGTAT  1120
NODE_68291_after  ATGATCTTCACTATGGGGAAATGACAGTTAGAGAGACATTGGACTTCTCGGGTCGCTGTTTGGGTGTAGGGACAAGGTAT   981

                  ********************************************************************************
NODE_68291_before GAGATGCTTGCAGAACTCTCCAGGCGAGAGAGGGAAGCCGGAATCAAACCTGATCCCGAAATTGACGCTTTTATGAAGGC  1200
NODE_68291_after  GAGATGCTTGCAGAACTCTCCAGGCGAGAGAGGGAAGCCGGAATCAAACCTGATCCCGAAATTGACGCTTTTATGAAGGC  1061

                  ********************************************************************************
NODE_68291_before CACAGCTCTGTCGGGTCAAGAGACAAGCTTGGTCACTGATTATATACTCAAGATTCTTGGATTGGATATCTGTGCAGACA  1280
NODE_68291_after  CACAGCTCTGTCGGGTCAAGAGACAAGCTTGGTCACTGATTATATACTCAAGATTCTTGGATTGGATATCTGTGCAGACA  1141

                  ********************************************************************************
NODE_68291_before TTATGGTCGGAGATGAGATGCGAAGGGGCATTTCAGGTGGACAAAAAAAGCGTCTTACAACCGGAGAGATGTTAGTAGGA  1360
NODE_68291_after  TTATGGTCGGAGATGAGATGCGAAGGGGCATTTCAGGTGGACAAAAAAAGCGTCTTACAACCGGAGAGATGTTAGTAGGA  1221

                  ********************************************************************************
NODE_68291_before CCAGCAAAGGCTCTTTTTATGGATGAAATATCCACAGGGTTGGACAGTTCCACTACTTTTCAAATTTGCAAATTCATGAG  1440
NODE_68291_after  CCAGCAAAGGCTCTTTTTATGGATGAAATATCCACAGGGTTGGACAGTTCCACTACTTTTCAAATTTGCAAATTCATGAG  1301

                  ********************************************************************************
NODE_68291_before GCAGATGGTTCATATTATGGATGTCACCATGATCATCTCATTGCTTCAGCCGGCTCCTGAGACTTATGATCTCTTCGATG  1520
NODE_68291_after  GCAGATGGTTCATATTATGGATGTCACCATGATCATCTCATTGCTTCAGCCGGCTCCTGAGACTTATGATCTCTTCGATG  1381
```

```
                  ***********************                            *****************************
NODE_68291_before ACATTATCCTTCTCTCGGAGGGT                            ACGTCCTCGAGTTTTTCGAGCACATGGGA  1600
NODE_68291_after  ACATTATCCTTCTCTCGGAGGGTCAAGTCGTCTACCAAGGTCCACGAGAGAACGTCCTCGAGTTTTTCGAGCACATGGGA  1461

                  ********************************************************************************
NODE_68291_before TTCAAGTGCCCTGAAAGGAAAGGAGTTGCCGACTTCTTGCAAGAAGTGACATCTAAGAAAGATCAAGAACAGTATTGGTT  1680
NODE_68291_after  TTCAAGTGCCCTGAAAGGAAAGGAGTTGCCGACTTCTTGCAAGAAGTGACATCTAAGAAAGATCAAGAACAGTATTGGTT  1541

                  ********************************************************************************
NODE_68291_before CAAGAAGAACCAACCTTTCCAATACGTTTCTGTAGATGATTTCGTGCATGGATTCAAATCTTTTCACATTGGCCAACATC  1760
NODE_68291_after  CAAGAAGAACCAACCTTTCCAATACGTTTCTGTAGATGATTTCGTGCATGGATTCAAATCTTTTCACATTGGCCAACATC  1621

                  ******************************************************************************
NODE_68291_before TGTCATCCGATCTTAGGATTCCTTATGACAAATCAAAAACTCACCCAGCTGCACTAGTCAAAGAGAAATACGGGNNNN--  1838
NODE_68291_after  TGTCATCCGATCTTAGGATTCCTTATGACAAATCAAAAACTCACCCAGCTGCACTAGTCAAAGAGAAATACGGGNNNNGC  1701


NODE_68291_before --------------------------------------------------------------------------------  1838
NODE_68291_after  ACTAGTCAAAGAGAAATACGGGATTTCAAATATGGAGCTGTTCAAGGCATGCTTTGCCAGAGAATGGCTACTAATGAAGC  1781

                       ***************************************************************************
NODE_68291_before -----TCCTTTGTTTACATATTCAAGACCACCCAGATCACTATCATGTCGCTTATTGCTCTGACGGTGTTCCTTAGGACT  1913
NODE_68291_after  GAAACTCCTTTGTTTACATATTCAAGACCACCCAGATCACTATCATGTCGCTTATTGCTCTGACGGTGTTCCTTAGGACT  1861

                  ********************************************************************************
NODE_68291_before GAAATGCCAGTAGGGTCAGTGCAAGATGGAGGGAAGTTTTTTGGAGCACTTTTCTTCAGCTTGATCAATGTCATGTTCAA  1993
NODE_68291_after  GAAATGCCAGTAGGGTCAGTGCAAGATGGAGGGAAGTTTTTTGGAGCACTTTTCTTCAGCTTGATCAATGTCATGTTCAA  1941

                  ********************************************************************************
NODE_68291_before TGGAATGGCGGAACTTGCAATGACCGTTTTCAGGCTTCCTGTGTTCTATAAGCAGAGAGATTTCTTGTTTTACCCCGCTT  2073
NODE_68291_after  TGGAATGGCGGAACTTGCAATGACCGTTTTCAGGCTTCCTGTGTTCTATAAGCAGAGAGATTTCTTGTTTTACCCCGCTT  2021

                  ********************************************************************************
NODE_68291_before GGGCTTTCGGCTTGCCTATTTGGGTCCTCCGAATTCCGTTGTCATTCATGGAATCAGGGATATGGATCATCTTAACATAC  2153
NODE_68291_after  GGGCTTTCGGCTTGCCTATTTGGGTCCTCCGAATTCCGTTGTCATTCATGGAATCAGGGATATGGATCATCTTAACATAC  2101

                  ********************************************************************************
NODE_68291_before TACACCATTGGCTTCGCTCCAGCGGCCAGCAGGTTCTTCAAGCAATTCTTGGCATTCTTTGGCATCCATCAGATGGCACT  2233
NODE_68291_after  TACACCATTGGCTTCGCTCCAGCGGCCAGCAGGTTCTTCAAGCAATTCTTGGCATTCTTTGGCATCCATCAGATGGCACT  2181

                  ********************************************************************************
NODE_68291_before GTCCCTCTTTCGGTTCATTGCTGCAGTTGGGAGAACTCAGGTTGTCGCAAACACCCTGGGAACCTTCACTTTGCTAATGG  2313
NODE_68291_after  GTCCCTCTTTCGGTTCATTGCTGCAGTTGGGAGAACTCAGGTTGTCGCAAACACCCTGGGAACCTTCACTTTGCTAATGG  2261

                  ********************************************************************************
NODE_68291_before TTTTCGTTCTTGGAGGATTTATTGTTTCCAAAAACGACATCGAGCCATGGATGATATGGGGATATTACGTATCTCCTATG  2393
NODE_68291_after  TTTTCGTTCTTGGAGGATTTATTGTTTCCAAAAACGACATCGAGCCATGGATGATATGGGGATATTACGTATCTCCTATG  2341

                  ********************************************************************************
NODE_68291_before ATGTATGGGCAAAATGCTATAGTGATGAATGAATTCCTCGACAAAAGATGGAGCACGCGTAACGAGGATACTAGAATTAA  2473
NODE_68291_after  ATGTATGGGCAAAATGCTATAGTGATGAATGAATTCCTCGACAAAAGATGGAGCACGCGTAACGAGGATACTAGAATTAA  2421

                  ********************************************************************************
NODE_68291_before TGAGCCCACAGTTGGAAAAGTGCTTTTGAAGTCTCGAGGTTTCTTCGTACAAGAATATTGGTATTGGATCTGCATTGGAG  2553
NODE_68291_after  TGAGCCCACAGTTGGAAAAGTGCTTTTGAAGTCTCGAGGTTTCTTCGTACAAGAATATTGGTATTGGATCTGCATTGGAG  2501

                  ********************************************************************************
NODE_68291_before CACTGTTTGGGTTTTCACTCCTCTTCAACATCTTGTTTGTTGCAGCATTGACTTGGTTAAATCCTTTGGGAGATGCAAAA  2633
NODE_68291_after  CACTGTTTGGGTTTTCACTCCTCTTCAACATCTTGTTTGTTGCAGCATTGACTTGGTTAAATCCTTTGGGAGATGCAAAA  2581

                  ********************************************************************************
NODE_68291_before GCAGTTGTCTCGGATGAAGAGGCGGATAAGAAGAAAAACAAATCATTGTCTTCGCAACTTGCGAAAGAAGGAATCGACAT  2713
NODE_68291_after  GCAGTTGTCTCGGATGAAGAGGCGGATAAGAAGAAAAACAAATCATTGTCTTCGCAACTTGCGAAAGAAGGAATCGACAT  2661

                  ********************************************************************************
NODE_68291_before GCAAGTGAGAAGTTCTTCTGAAATCGTTAGCACTTCAGAGAATATACAGAGAAGAGGGATGGTTCTGCCATTCCAACCCC  2793
NODE_68291_after  GCAAGTGAGAAGTTCTTCTGAAATCGTTAGCACTTCAGAGAATATACAGAGAAGAGGGATGGTTCTGCCATTCCAACCCC  2741

                  ********************************************************************************
NODE_68291_before TTTCTCTTGCGTTCAACCATGTGAACTACTACGTGGATATGCCTGCAGAAATGAAGAGTCAAGGAGTTGAGGAAGACCGT  2873
NODE_68291_after  TTTCTCTTGCGTTCAACCATGTGAACTACTACGTGGATATGCCTGCAGAAATGAAGAGTCAAGGAGTTGAGGAAGACCGT  2821

                  ********************************************************************************
NODE_68291_before CTCCAACTGTTGAGAGATGTCAGTGGCGCTTTCAGACCAGGGGTACTCACAGCATTGGTCGGGGTTAGTGGTGCTGGAAA  2953
NODE_68291_after  CTCCAACTGTTGAGAGATGTCAGTGGCGCTTTCAGACCAGGGGTACTCACAGCATTGGTCGGGGTTAGTGGTGCTGGAAA  2901
```

```
                  *********************** [illegible] *******************************
NODE_68291_before GACAACCCTCATGGATGTGCTAG [illegible] AGGAAGTATTAGCATCTCCGGATACCCTA   3033
NODE_68291_after  GACAACCCTCATGGATGTGCTAGCAGGAAGGAAGACAGGTGGTTACATAGAAGGAAGTATTAGCATCTCCGGATACCCTA   2981

                  ********************************************************************************
NODE_68291_before AAAACCAATCAACGTTTGCTCGGGTCAGTGGTTACTGTGAACAGAACGACATTCACTCGCCTAACGTCACTGTCTACGAA   3113
NODE_68291_after  AAAACCAATCAACGTTTGCTCGGGTCAGTGGTTACTGTGAACAGAACGACATTCACTCGCCTAACGTCACTGTCTACGAA   3061

                  ********************************************************************************
NODE_68291_before TCCCTCCTATACTCAGCCTGGCTTCGTCTTTCTTCCGACATTAAGACTCAAACTCGCAAGATGTTTGTGGAAGAAGTTAT   3193
NODE_68291_after  TCCCTCCTATACTCAGCCTGGCTTCGTCTTTCTTCCGACATTAAGACTCAAACTCGCAAGATGTTTGTGGAAGAAGTTAT   3141

                  ********************************************************************************
NODE_68291_before GGAGTTGGTTGAGCTCAACCCTATCAGAAACGCGCTTGTCGGGCTTCCTGGTGTCGATGGCCTTTCGACTGAGCAAAGAA   3273
NODE_68291_after  GGAGTTGGTTGAGCTCAACCCTATCAGAAACGCGCTTGTCGGGCTTCCTGGTGTCGATGGCCTTTCGACTGAGCAAAGAA   3221

                  ********************************************************************************
NODE_68291_before AGCGGCTGACAATAGCTGTAGAGTTGGTGGCTAATCCATCTATTATCTTTATGGACGAACCAACCTCCGGCCTTGATGCT   3353
NODE_68291_after  AGCGGCTGACAATAGCTGTAGAGTTGGTGGCTAATCCATCTATTATCTTTATGGACGAACCAACCTCCGGCCTTGATGCT   3301

                  ********************************************************************************
NODE_68291_before AGAGCAGCCGCCATCGTGATGCGTACGGTGAGGAACACGGTGGATACAGGGAGGACTGTTGTTTGCACGATTCACCAGCC   3433
NODE_68291_after  AGAGCAGCCGCCATCGTGATGCGTACGGTGAGGAACACGGTGGATACAGGGAGGACTGTTGTTTGCACGATTCACCAGCC   3381

                  ********************************************************************************
NODE_68291_before GAGCATTGACATTTTTGAAGCTTTTGATGAGTTGCTATTAATGAAAAGAGGCGGGCGGGTCATTTATGCTGGCCCTCTTG   3513
NODE_68291_after  GAGCATTGACATTTTTGAAGCTTTTGATGAGTTGCTATTAATGAAAAGAGGCGGGCGGGTCATTTATGCTGGCCCTCTTG   3461

                  ********************************************************************************
NODE_68291_before GTCGCCATTCCCACAAGCTCGTAGAATATTTTGAGGCTGTCCCAGGGGTTCCGAAGATCAGGGATGGTCACAATCCAGCC   3593
NODE_68291_after  GTCGCCATTCCCACAAGCTCGTAGAATATTTTGAGGCTGTCCCAGGGGTTCCGAAGATCAGGGATGGTCACAATCCAGCC   3541

                  ********************************************************************************
NODE_68291_before ACATGGATGCTTGAAGTGAGTGCTCCGGCAGTTGAGGCTCAGCTCGAGGTCGACTTCGCAGATATTTACCCAAACTCTGA   3673
NODE_68291_after  ACATGGATGCTTGAAGTGAGTGCTCCGGCAGTTGAGGCTCAGCTCGAGGTCGACTTCGCAGATATTTACCCAAACTCTGA   3621

                  ********************************************************************************
NODE_68291_before CCTTTATAAGCGGAACCAAGACCTGATCAAAGAGCTTAGTACCCCAGCCCCAGGCTGCAAAGATCTCCACTTCCCTACCG   3753
NODE_68291_after  CCTTTATAAGCGGAACCAAGACCTGATCAAAGAGCTTAGTACCCCAGCCCCAGGCTGCAAAGATCTCCACTTCCCTACCG   3701

                  ********************************************************************************
NODE_68291_before AGTACTCACAACCTTTCCTCACTCAGTGCAAGGCTTGTTTCTGGAAACAGCACTGGTCTTACTGGAGAAATCCTCAGTAC   3833
NODE_68291_after  AGTACTCACAACCTTTCCTCACTCAGTGCAAGGCTTGTTTCTGGAAACAGCACTGGTCTTACTGGAGAAATCCTCAGTAC   3781

                  ********************************************************************************
NODE_68291_before AACGCCATCCGGTTCTTTATGACCATAGTCATCGCCATTTTGTTTGGTTTAATATTCTGGGATAAAGGACAGCAGACGAC   3913
NODE_68291_after  AACGCCATCCGGTTCTTTATGACCATAGTCATCGCCATTTTGTTTGGTTTAATATTCTGGGATAAAGGACAGCAGACGAC   3861

                  ********************************************************************************
NODE_68291_before CAAGCAACAAGACCTGATGAATCTTTTGGGAGCCATGTACGCAGCTGTGCTTTTCCTTGGGGCCACAAATGCTTCTGCTG   3993
NODE_68291_after  CAAGCAACAAGACCTGATGAATCTTTTGGGAGCCATGTACGCAGCTGTGCTTTTCCTTGGGGCCACAAATGCTTCTGCTG   3941

                  ********************************************************************************
NODE_68291_before TGCAGTCTATAGTCGCCATTGAGAGGACAGTCTTCTACCGTGAACGAGCAGCTGGAATGTACTCTCCGCTGCCATACGCA   4073
NODE_68291_after  TGCAGTCTATAGTCGCCATTGAGAGGACAGTCTTCTACCGTGAACGAGCAGCTGGAATGTACTCTCCGCTGCCATACGCA   4021

                  ********************************************************************************
NODE_68291_before TTTGCTCAGGTGGCTATTGAGACAATTTATGTAGCGATTCAGACATTGGTCTACAGTCTTCTCCTTTACTCGATGATTGG   4153
NODE_68291_after  TTTGCTCAGGTGGCTATTGAGACAATTTATGTAGCGATTCAGACATTGGTCTACAGTCTTCTCCTTTACTCGATGATTGG   4101

                  ********************************************************************************
NODE_68291_before GTTCGAGTGGAAGGCGGGGAAGTTCTTGTGGTTCTACTACTACATACTGATGTGCTTCATCTACTTCACGATGTATGGAA   4233
NODE_68291_after  GTTCGAGTGGAAGGCGGGGAAGTTCTTGTGGTTCTACTACTACATACTGATGTGCTTCATCTACTTCACGATGTATGGAA   4181

                  ********************************************************************************
NODE_68291_before TGATGGTTGTAGCATTGACACCAGGCCACCAGATAGCTGCCATTGTGATGTCCTTCTTCCTGAGCTTCTGGAACTTGTTC   4313
NODE_68291_after  TGATGGTTGTAGCATTGACACCAGGCCACCAGATAGCTGCCATTGTGATGTCCTTCTTCCTGAGCTTCTGGAACTTGTTC   4261

                  ********************************************************************************
NODE_68291_before TCTGGCTTCCTTATCCCTAGGCCGCAAATTCCTGTATGGTGGAGGTGGTATTACTGGGCTTCACCAGTGGCATGGACGCT   4393
NODE_68291_after  TCTGGCTTCCTTATCCCTAGGCCGCAAATTCCTGTATGGTGGAGGTGGTATTACTGGGCTTCACCAGTGGCATGGACGCT   4341

                  ********************************************************************************
NODE_68291_before GTACGGTCTTGTCACCTCTCAAGTGGGCGACAAGAATGGCAATCTCGAAATACCAGGAGCCGGCAACATGCCGTTGAAGC   4473
NODE_68291_after  GTACGGTCTTGTCACCTCTCAAGTGGGCGACAAGAATGGCAATCTCGAAATACCAGGAGCCGGCAACATGCCGTTGAAGC   4421
```

```
                  ********************************************************************************
NODE_68291_before AGTTCCTGAAGGTAGAACTGGGTTTTGACTACAGCTTCCTCCCCGCTGTCGCGGTTGCTCACATCGGCTGGGTCCTTCTC  4553
NODE_68291_after  AGTTCCTGAAGGTAGAACTGGGTTTTGACTACAGCTTCCTCCCCGCTGTCGCGGTTGCTCACATCGGCTGGGTCCTTCTC  4501

                  ********************************************************************************
NODE_68291_before TTTTTCTTTGTCTTCGCTTACGGCATCAAGTTCCTCAATTTCCAGAGGAGATAAAACCGATGGCAAACAGTTCTCACTTT  4633
NODE_68291_after  TTTTTCTTTGTCTTCGCTTACGGCATCAAGTTCCTCAATTTCCAGAGGAGATAAAACCGATGGCAAACAGTTCTCACTTT  4581

                  ********************************************************************************
NODE_68291_before CTGGCTAGATTTTGAAACGTTAAACGTAGGCCCATCATGTAAATTAAGGATGATAGGCGACTAAAGAGTCTCCCTCCTCC  4713
NODE_68291_after  CTGGCTAGATTTTGAAACGTTAAACGTAGGCCCATCATGTAAATTAAGGATGATAGGCGACTAAAGAGTCTCCCTCCTCC  4661

                  ************************************************************************* ******
NODE_68291_before TGTTTTCTTCACTTTTCAGTAAGTCTTGCTTTTGTAACACTAGCATTCTTTGTCACCGCTGCTTCATTGGACTGAGAGCG  4793
NODE_68291_after  TGTTTTCTTCACTTTTCAGTAAGTCTTGCTTTTGTAACACTAGCATTCTTTGTCACCGCTGCTTCATTGGACTTAGAGCG  4741

                  ** *******
NODE_68291_before TCAGTTAATTGTAAGAGACAAATAATTAATTTGAAATGCAAACGAGTGGTGTG  4846
NODE_68291_after  TCGGTTAATT-------------------------------------------  4751
```

**Appendix D**

# *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq

Mizrachi et al. BMC Genomics 2010, **11**:681
http://www.biomedcentral.com/1471-2164/11/681

BMC Genomics

**RESEARCH ARTICLE** **Open Access**

# *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq

Eshchar Mizrachi[1†], Charles A Hefer[2†], Martin Ranik[1], Fourie Joubert[2], Alexander A Myburg[1*]

## Abstract

**Background:** *De novo* assembly of transcript sequences produced by short-read DNA sequencing technologies offers a rapid approach to obtain expressed gene catalogs for non-model organisms. A draft genome sequence will be produced in 2010 for a *Eucalyptus* tree species (*E. grandis*) representing the most important hardwood fibre crop in the world. Genome annotation of this valuable woody plant and genetic dissection of its superior growth and productivity will be greatly facilitated by the availability of a comprehensive collection of expressed gene sequences from multiple tissues and organs.

**Results:** We present an extensive expressed gene catalog for a commercially grown *E. grandis* × *E. urophylla* hybrid clone constructed using only Illumina mRNA-Seq technology and *de novo* assembly. A total of 18,894 transcript-derived contigs, a large proportion of which represent full-length protein coding genes were assembled and annotated. Analysis of assembly quality, length and diversity show that this dataset represent the most comprehensive expressed gene catalog for any *Eucalyptus* tree. mRNA-Seq analysis furthermore allowed digital expression profiling of all of the assembled transcripts across diverse xylogenic and non-xylogenic tissues, which is invaluable for ascribing putative gene functions.

**Conclusions:** *De novo* assembly of Illumina mRNA-Seq reads is an efficient approach for transcriptome sequencing and profiling in *Eucalyptus* and other non-model organisms. The transcriptome resource (Eucspresso, http://eucspresso.bi.up.ac.za/) generated by this study will be of value for genomic analysis of woody biomass production in *Eucalyptus* and for comparative genomic analysis of growth and development in woody and herbaceous plants.

## Background

Ultra-high-throughput second-generation DNA sequencing technologies from companies such as Roche (454 pyrosequencing), Illumina (sequencing by synthesis, Solexa GA) and Applied Biosystems (sequencing by ligation, SOLiD), are increasingly being used for novel exploratory genomics in small to medium-sized laboratories. "Short-read" (36 - 72 nt) technologies such as those of Illumina and Applied Biosystems have proven to be exceptionally successful in a wide variety of whole-transcriptome investigations [1-5], but most of these studies have relied on prior sequence knowledge such as an annotated genome for qualitative and quantitative transcriptome analyses.

Genome assembly of short sequences without any auxiliary knowledge has primarily utilized 454 sequencing data, due to the longer individual read lengths of 150-400 base pairs (bp). However, short-read sequencing (Illumina GA and SOLiD) has been successfully used for *de novo* assembly of small bacterial genomes (2-5 Mbp), where 36 bp reads have been assembled [6-8] and hybrid approaches, where genomes are *de novo* assembled using a combination of reads from multiple sequencing platforms to overcome the inherent limitations of each technology, have been used to successfully assemble genomes of up to 40 Mbp [9,10]. More recently, the sequencing of the giant panda genome was demonstrated [11] using *de novo* assembly of sequence derived from a single platform (Illumina), but utilizing a

---

* Correspondence: zander.myburg@fabi.up.ac.za
† Contributed equally
[1] Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa
Full list of author information is available at the end of the article

combination of different insert sizes, allowing assembly of an estimated 94% of the genome (2.25 Gbp). *De novo* assembly of large, highly repetitive and highly heterozygous eukaryotic genomes from short-read data remains a challenge.

In transcriptome studies, 454 pyrosequencing has proven very useful for generating ESTs representing the majority of expressed genes. This has enabled gene discovery in a variety of previously uncharacterized eukaryotic organisms with no or little *a priori* DNA sequence information [12-16]. However, relatively few published studies have attempted *de novo* assembly of whole-transcriptome sequences from short-read data such as that generated by Illumina GA or SOLiD technologies. Assembly of short (36-72 bp) read data into accurate, contiguous transcript sequences has only recently been reported [17-19] demonstrating that assembly of long, potentially full-length, transcript assemblies is indeed possible.

*Eucalyptus* tree species and hybrids presently constitute the most widely planted (≈ 20 Mha) and commercially important hardwood fibre crop in the world. They are mainly utilized for timber, pulp and paper production [20]. Their fast growth rates and wide adaptability may in future allow sustainable and cost efficient production of woody biomass for bioenergy generation [21,22]. *Eucalyptus* will soon be only the second forest plantation genus (after *Populus*) for which a reference genome sequence will be completed by end 2010 [23]. To support the genome annotation effort, there is much value in having a dataset of genes with strong transcriptional evidence across a range of tissues and developmental stages. Until recently, limited amounts of *Eucalyptus* EST/unigene data were available in public databases, mainly due to the fact that commercial interests have necessitated private EST collections [24]. As of March 2010, aside from a mixed-species collection of ≈56,000 nucleotide sequences on NCBI (≈ 37,000 of which are Sanger EST sequences) and which contain extensive redundancy, the largest effort to date to generate a comprehensive catalogue of expressed genes in a single *Eucalyptus* species was based on 454 sequencing of cDNA fragments from *E. grandis* trees [15]. While this study provided an excellent representation of expressed genes and gene ontology classes in *E. grandis*, the relatively short lengths of the assembled contigs (mean length of 389 bp for all contigs longer than 200 bp) meant that very few complete gene models were represented. There remains therefore a fundamental need for a high-quality expressed gene catalog for *Eucalyptus*, to support genome annotation efforts and discern authentically expressed genes from predicted gene models, as well as for future genomics research, which will include transcriptome, proteome and metabolome profiling.

In the process of producing such a high-quality expressed gene catalog for *Eucalyptus*, we addressed three main questions: First, is it feasible to *de novo* assemble Illumina mRNA-Seq data into contiguous, near full-length gene model sequences for *Eucalyptus*? Second, what genes make up the expressed gene catalog for a fast-growing *Eucalyptus* plantation tree? Finally, can we re-use the mRNA-Seq data to create a tissue and organ-specific digital expression profile for each assembled contig? We addressed these questions by generating a comprehensive set of expressed gene sequences from a commercially grown *Eucalyptus* hybrid (*E. grandis* × *E. urophylla*) clone using Illumina mRNA-Seq technology and *de novo* short-read assembly. We report herein the complete annotation of the expressed gene catalog based on comparative analysis with the published *Arabidopsis thaliana* [25], *Populus trichocarpa* [26] and *Vitis vinifera* [27] protein-coding datasets. We describe an interactive database of annotated transcript sequences, coding sequences (CDSs) and derived protein sequences (Eucspresso, http://eucspresso.bi.up.ac.za/, CA Hefer, E Mizrachi, AA Myburg, F Joubert, unpublished), which will be continuously updated and curated in association with the *Eucalyptus* Genome Network (EUCAGEN, http://www.eucagen.org) as part of an effort to initiate a publicly accessible database for *Eucalyptus* transcriptomics research similar to that produced for *Populus* [28].

## Results

### *De novo* assembly, validation and annotation of contigs

In total, 62 million paired-end reads of raw mRNA-Seq data (6.90 Gbp) representing poly(A)-selected RNA from six *Eucalyptus* tissues and varying in lengths from 36 bp to 60 bp, were generated in 14 lanes on Illumina GA and GAII instruments. Following a sequence filtering process to exclude low quality and ribosomal RNA-derived reads, we assembled 36 million paired-end reads (3.93 Gbp, Additional file 1 - Table S1 and Figure S1, NCBI Sequence Read Archive accession SRA012408) of non-normalized mRNA sequence, using the Velvet short-read assembler (version 0.7.30, [29]). In total, 18,894 RNA-derived contigs were assembled (comprising 22.1 Mbp of transcriptome sequence) that were greater than 200 bp in length (mean = 1170 bp, Figure 1 and Additional file 2), with a median coverage per base (CPB) per contig of 37×, ranging from 8× (minimum coverage cut-off for assembly) to 5,262× (Additional file 1 Figure S2).

We performed *ab inito* CDS prediction using GENSCAN [30] and found that 15,713 contigs (83.2%) contained a predicted CDS (Additional file 1 Table S3). Analysis of the predicted coding sequences using Anaconda [31] identified 6,208 contigs that contained

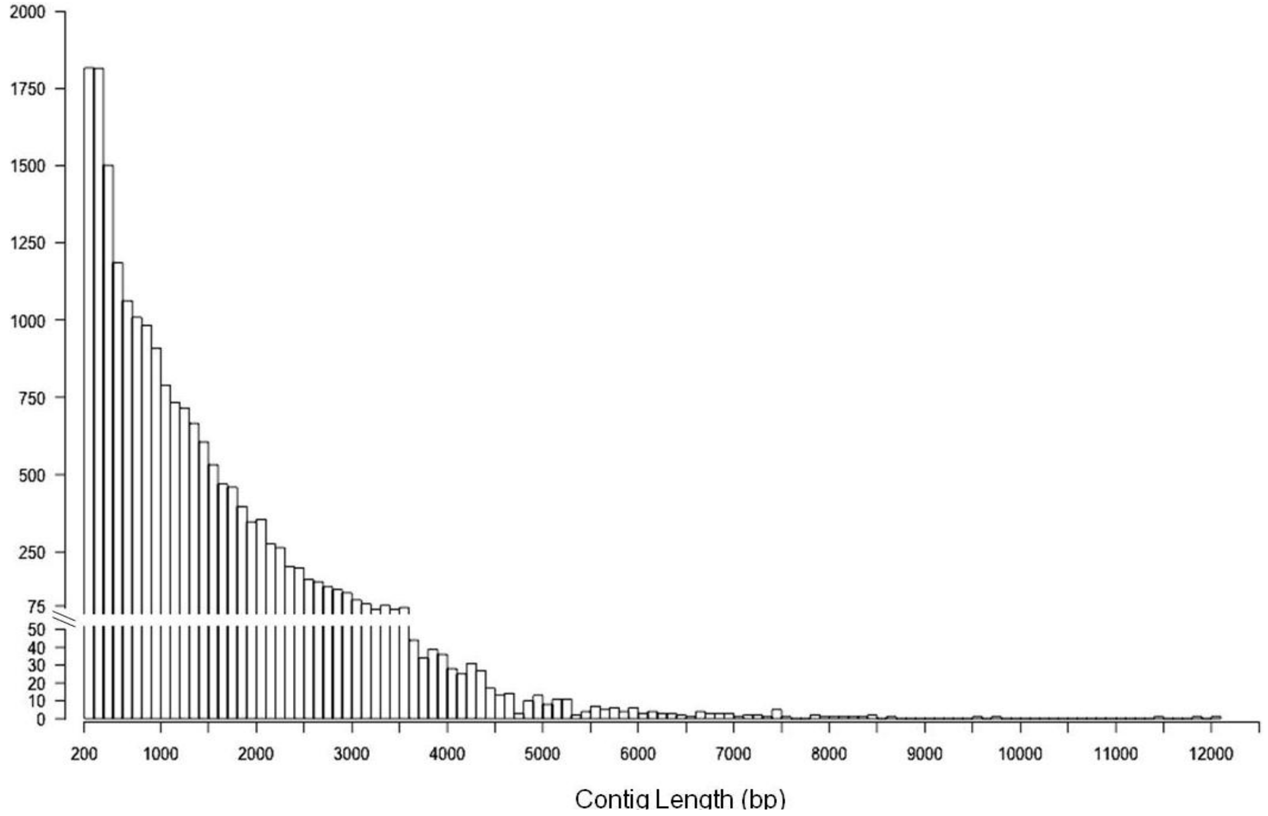


**Figure 1 Summary distribution of the lengths of the 18,894 assembled contigs (> 200 bp, mean length = 1170 bp, N50 = 1,640 bp, Q3 = 1,573 bp, Max = 12,053 bp)**.

putatively full-length CDSs (i.e. containing start and stop codons), 4,610 predicted to contain a start but no stop codon, 4,874 predicted to contain a stop but no start codon, and only 21 with neither. To ascertain the quality of Velvet assembly of short reads into long contiguous coding sequences, we compared a subset of 35 of our transcript-derived contigs to corresponding Sanger-sequenced, full-length, cloned *Eucalyptus grandis* mRNA sequences in NCBI (Figure 2 and Additional file 3). Paired reads were independently mapped to each Sanger reference sequence, the *de novo* assembled Velvet contig and its corresponding predicted CDS. A Needleman-Wunsch alignment of these three sequences was used for contiguity validation of the assembled contigs. Independently, each sequence had 100% coverage validation across the contig, except in cases of low quality assembly ('N's inserted by Velvet), which occurred in regions of coverage lower than 8× per base. Of the 35 transcript-derived contigs evaluated, 25 (71%) assembled completely with a 5' UTR, 3' UTR, as well as a contiguous coding sequence matching that of the reference mRNA sequence. We found several cases where, despite high coverage, our transcript-derived contigs differed from the Sanger reference sequence due to indels, but these were generally in the UTR regions and likely represent allelic differences between the F1 hybrid individual and the reference sequences (Additional file 3).

Of the 18,894 assembled contigs, 18,606 (98.48%) exhibited significant similarity (BLASTN, -10, [31]) to the preliminary draft 8X DOE-JGI *E. grandis* genome assembly (http://eucalyptusdb.bi.up.ac.za/) consistent with the origin of the mRNA contigs (an F1 hybrid of *E. grandis* and *E. urophylla*). We further characterized the assembled contigs by high stringency BLASTX analysis (-10 confidence, minimum 100 bp high scoring pair (HSP) match length) to protein datasets from three reference sequenced angiosperm genera (*Arabidopsis*, *Populus* and *Vitis*). Cumulatively, 15,055 contigs (79.68%) exhibited high similarity to *Arabidopsis* (14,235 contigs), *Populus* (14,769 contigs) or *Vitis* proteins (14,833 contigs, Additional file 1 Figure S3). Of the 15,055 contigs with high similarity to *Arabidopsis*, *Populus* or *Vitis* proteins, 13,806 (91.70%) also contained predicted coding sequences (Figure 3A), while 1,249 (8.30%) did not (Figure 3B), possibly due to low expression of these transcripts which would have resulted in lower coverage and shorter contigs that represented only a fraction of the open reading frame (or mostly

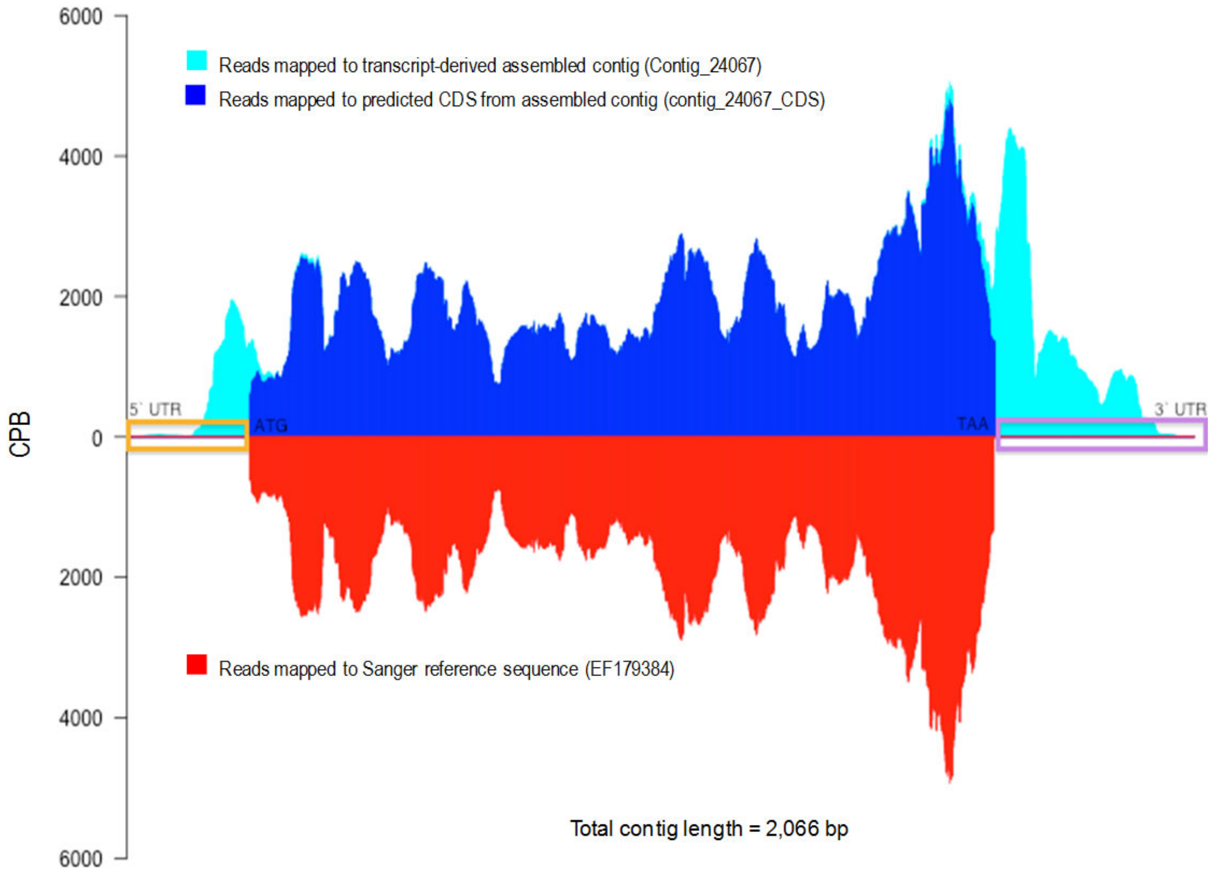


**Figure 2 Comparison of the *de novo* assembled contig of the *Eucalyptus grandis* UDP-glucose dehydrogenase (*UGDH*) transcript to a reference Sanger-based sequence (Genbank** EF179384**) for the same gene**. Peak height indicates coverage per base (CPB) of mapped short-reads across each sequence. CPB of the fully assembled contig is shown in cyan. CPB of the predicted CDS region is shown in dark blue. CPB of the Sanger reference sequence is shown in red. 5′ UTR (orange box) and 3′ UTR (purple box) regions are indicated.

UTR sequence). Predicted codon usage and amino acid frequencies in the proteome represented by the *Eucalyptus* expressed gene catalog were very similar to those of expressed gene catalogs from *Arabidopsis* and *Populus* (Additional file 1 Figure S4 and Figure S5).

To compare the completeness of our expressed gene catalogue to that of all publicly available gene sequence data for *Eucalyptus*, we generated a separate dataset, termed EucALL, containing all publicly available *Eucalyptus* gene sequence data to date (March 2010). This included all NCBI unigenes and ESTs, assembled 454 EST data from *E. grandis* leaf tissue (DOE-JGI, http://eucalyptusdb.bi.up.ac.za/), assembled 454 EST data produced by Novaes and colleagues [15], and the EucaWood contig dataset [33]. We compared the representation of *Arabidopsis* genes in the EucALL dataset and in our assembled *E. grandis* × *E. urophylla* (EGU) transcript dataset by BLASTX at significance levels of < $1e^{-05}$, < $1e^{-10}$ and < $1e^{-20}$ (Additional file 1 Table S2). While the overall numbers of hits were higher in the EucALL dataset, these were mostly in the lower size ranges. For our *de novo* assembled contigs, a much higher number of significant hits in contigs larger than 2000 bp in size (6,602 compared to 1,940 at significance < $1e^{-10}$) suggested that a greater proportion of our contigs represent full-length gene models than the publicly available *Eucalyptus* gene sequence set (EucALL).

### Functional annotation of the expressed gene catalog

The transcript-derived contig sequences were annotated according to several functional annotation conventions, including Gene Ontology (GO - http://www.geneontology.org/), KEGG (http://www.genome.jp/kegg/) and InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/). The numbers and assortment of allocated GO categories provides a good indication of the large diversity of expressed genes sampled from the *Eucalyptus* transcriptome (Figure 4). This was also reflected in the diversity of InterProScan categories identified (Additional file 1 Figure S6 and Figure S7), as well as the

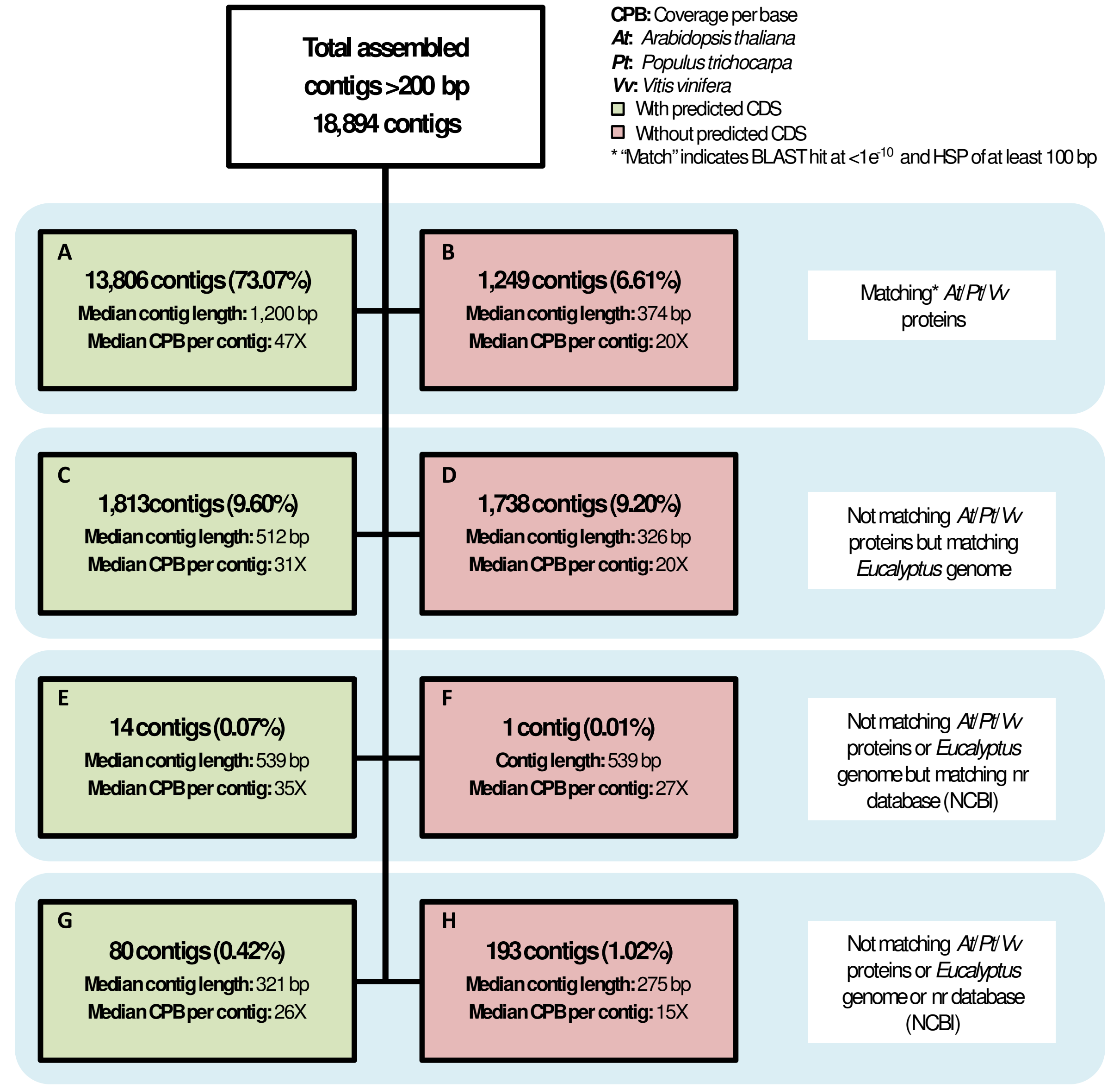


**Figure 3 Breakdown of annotation categories for all 18,894 transcript-derived contigs**. A large proportion (98.5%) of assembled contigs (A-D) had significant BLAST hits (< $1e^{-10}$ confidence, minimum 100 bp HSP match length) to the draft *Eucalyptus* genome assembly (http://eucalyptusdb.bi.up.ac.za/), 80% of which (A, B) also exhibited significant similarity (BLASTX, < $1e^{-10}$, > 100 bp HSP) to coding sequences of *Arabidopsis*, *Populus* or *Vitis*.

comprehensive coverage of biochemical processes by KEGG annotation, which was similar to that of the entire *Arabidopsis* gene catalog (Additional file 1 Figure S8).

## Digital expression profiling

An accepted method of identifying large scale differences in gene expression is to use EST abundance as an indicator of transcript abundance. This method has been implemented and validated in numerous studies using Sanger-derived ESTs [34,35], as well as 454-pyrosequencing methods [13,36-39]. Quantitative transcriptome analysis using ultra-high-throughput sequencing technologies such as Illumina and SOLiD has been shown to be accurate and highly correlated with other quantitative methods such as RT-qPCR and microarray analysis [1,5]. To quantify tissue-specific transcript abundance reflected in our short-read dataset, we combined data (multiple lanes in most cases) generated from the same tissues and mapped six tissue-specific datasets (Additional file 1 Table S1) to the assembled gene catalog using Bowtie [40]. Following this, we used

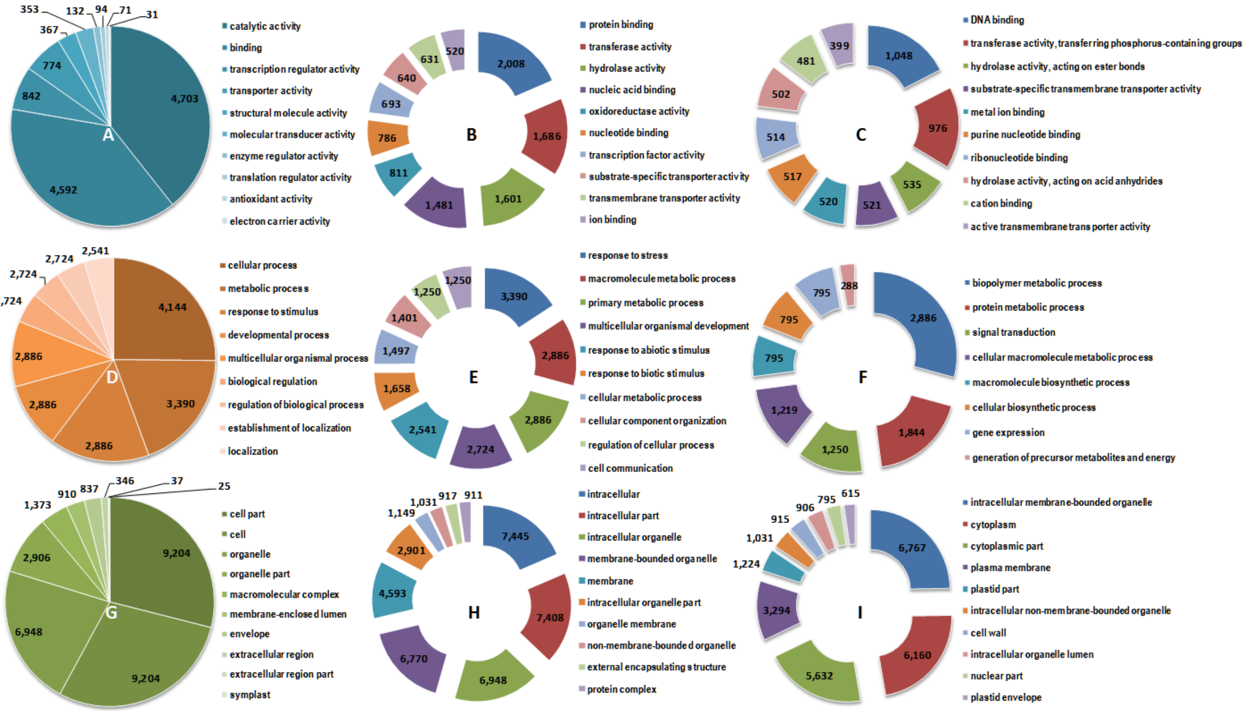


**Figure 4 Top ten most represented GO categories under the "Molecular Function" (A-C), "Biological Process" (D-F) and "Cellular Compartment" (G-I) categories in level 2 (A, D and G), 3 (B, E and H) and 4 (C, F and I).** The numbers and proportions in all categories reflect the diversity and complexity of genes expressed in multiple tissues sampled to make up the *Eucalyptus* gene catalog.

the Cufflinks [41] program (http://cufflinks.cbcb.umd.edu), which provides relative abundance values by calculating Fragments Per Kilobase of exon per Million fragments mapped (FPKM) as validated previously [2]. This enabled the allocation of a tentative digital expression profile for each transcript-derived contig (Additional file 4).

To compare between two general tissue types that are of interest for woody biomass production, we evaluated groups of genes whose FPKM values were greater than two-fold higher in woody (xylogenic) tissues (average FPKM of immature xylem and xylem: 1,897 annotated contigs) or leaf (non-xylogenic) tissues (average FPKM of shoot tips, young leaves and mature leaves: 1,531 annotated contigs). GO categories over-represented in the xylem-upregulated set compared to the leaf set (Figure 5A) was representative of developing woody tissues, with significant enrichment ($p < 0.05$) in signalling ("kinase activity"), carbohydrate metabolism, and genes associated with the Golgi, cytoskeleton and the plasma membrane - consistent with an emphasis on delivery of biopolymers to the cell wall. In contrast, gene categories significantly enriched ($p < 0.05$) in leaf tissue compared to woody tissue (Figure 5B) were associated with photosynthesis ("plastid", "thylakoid", "photosynthesis"), growth and energy production (precursor metabolites, "lipid biosynthesis", "amino acid metabolism").

We also interrogated our transcriptome data using the "core xylem gene set" identified in *Arabidopsis* by Ko and colleagues [42]. Of the 52 genes identified by the authors as markers of secondary xylem formation in *Arabidopsis*, 33 had putative homologues in the *Eucalyptus* transcriptome (BLASTX, < $1e^{-10}$) and in total 43 contigs were identified. Of these, 40 (93%) showed greater than two-fold "Xylem" to "Leaf" digital expression profile ratios and six were only detected in xylem tissues (Additional file 1 Table S4). Most of the expression profiles were also highly correlated with that of secondary cell wall-specific *Eucalyptus* cellulose synthase genes, similar to the patterns previously observed in *Arabidopsis*. These results are comparable to the 80% (51 out of 63 genes) reported recently for the same set of *Arabidopsis* homologs in *Populus* [43], which provided further support for the biological validity of the short-read-based digital expression profiles associated with the *Eucalyptus* expressed gene catalog.

## Public data resource

We constructed a public data resource, Eucspresso (http://eucspresso.bi.up.ac.za), which provides a

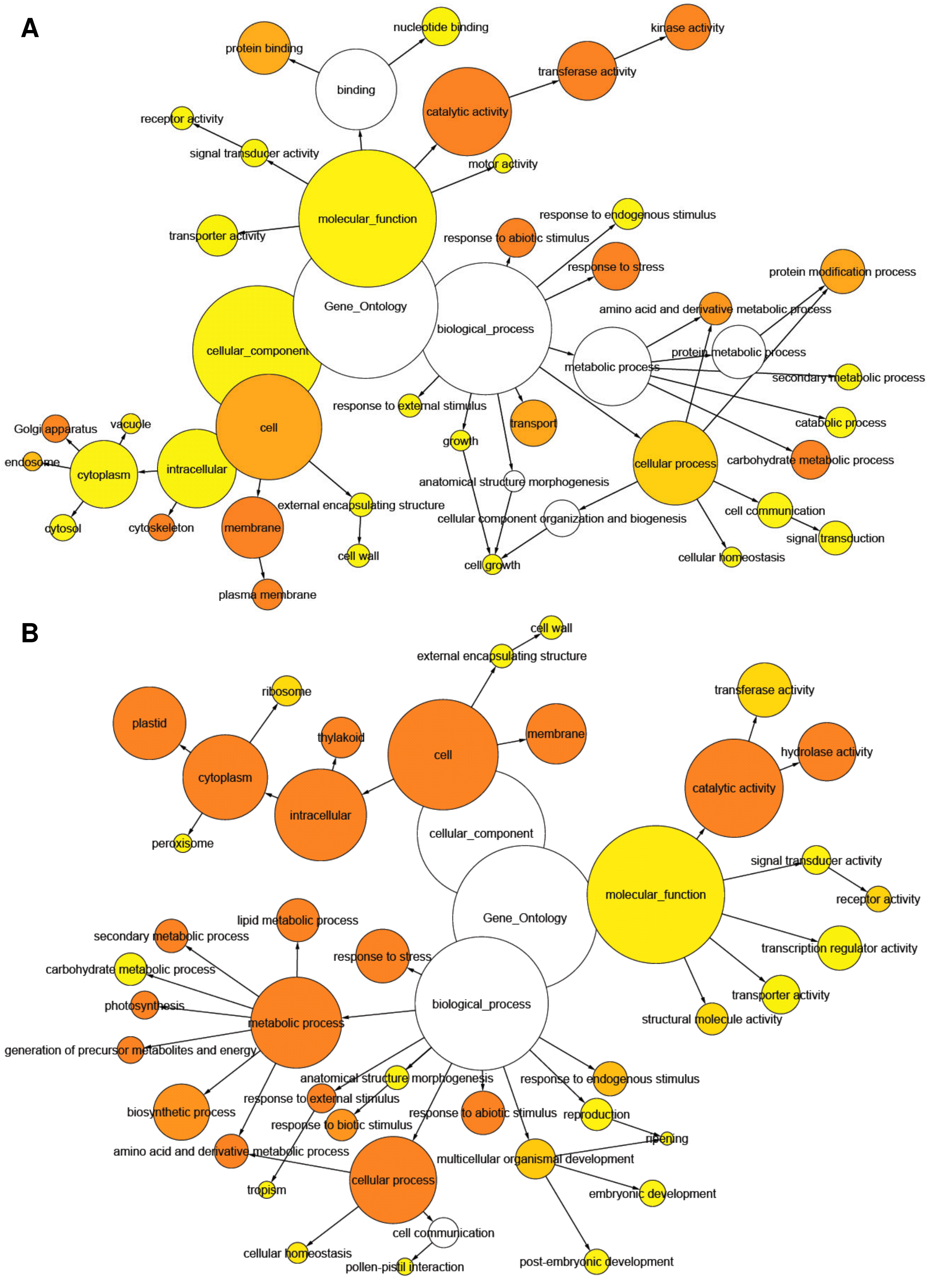


**Figure 5 Over-represented GO categories in xylem (A - 1,897 annotated contigs) and leaf (B - 1,531 annotated contigs) tissues**. All genes with a FPKM value more than two-fold higher in one tissue type versus the other were considered for the analysis. Data were analyzed using BiNGO (Maere et al. 2005). Node size is proportional to the number of genes in each category and colors shaded according to significance level (white - no significant difference, yellow - FDR = 0.05, Orange - FDR < 0.05).

searchable interface to the assembled contigs. The database can be queried based on closest homologous entry in the *Arabidopsis thaliana* (TAIR9), *Populus trichocarpa* (Version 2.0) and *Vitis vinifera* (Sept 2009 build) sequence data sets. Simple and compound keyword searches can be performed based on all of the functional annotation terms and the predicted coding and protein sequences can be obtained for all contigs. Finally, the tissue-specific (FPKM) digital expression profile and the location of each contig in the draft 8X *E. grandis* genome assembly (http://eucalyptusdb.bi.up.ac.za/) can be viewed from within Eucspresso.

## Discussion

We have assembled nearly 19,000 expressed gene sequences from xylogenic and non-xylogenic tissues of an actively growing *Eucalyptus* plantation tree using only Illumina mRNA-Seq technology and *de novo* short-read assembly. Quality control comparisons to full-length, cloned, Sanger-derived transcript sequences from *Eucalyptus*, as well as multiple lines of evidence such as CDS prediction and Pfam prediction showed that the transcript assemblies are robust and that thousands of full-length coding sequences and their respective 5' and/or 3' UTR regions were successfully assembled. Comparison of assembled gene models to gene catalogs of other angiosperm species by BLAST analysis and functional annotation (GO, InterProScan and KEGG category numbers and proportions, Figure 4 and Additional file 1 - Figure S6, Figure S7 and Figure S8) indicate that we have sampled an expansive and diverse expressed gene catalog representing a large proportion of the genes expressed in mature *Eucalyptus* trees across a variety of woody and non-woody tissues. Comparison to all publicly available *Eucalyptus* DNA sequence suggests that we have sampled a more comprehensive set of genes, which is also more complete in length (Additional file 1 - Table S2) from a single eucalypt tree genotype than has been available to date for the entire genus. Additionally, using a validated approach to quantify mRNA-Seq data we have produced an informative database of transcript abundance across six *Eucalyptus* tree tissues, which, due to the depth of sequencing, results in higher sensitivity and wider dynamic range than Sanger or 454-derived EST counts usually associated with this type of analysis.

A concern associated with *de novo* assembly of transcript sequences, be it Sanger derived [33] or 454 sequence derived [15] assemblies, is the contiguity of assembled sequences. This concern intuitively increases as the read length decreases, and may be one of the main reasons why most transcriptome *de novo* assembly approaches have utilized technologies with longer read lengths to date. We provide several lines of evidence which jointly support the contiguity of transcript sequences assembled in our study using short-read data. First, a high proportion of the contigs exhibited high-confidence BLASTX similarity to protein sequences from annotated gene catalogs of three angiosperm species *Arabidopsis*, *Populus* and *Vitis* (Figure 3). Second, a large proportion of the contigs contained long, near full-length, predicted CDSs (Figure 3). Third, InterproScan analysis predicted 45,687 protein domains, which is indicative of contiguous, in-frame predicted protein sequences (Additional File 1). Finally, a random subset of the contigs, which represented a variety of length and read coverage, were validated by direct alignment to previously published, Sanger sequenced, full-length *Eucalyptus* genes that were directly cloned from cDNA (Additional File 3).

Assigning biological significance to *de novo* assembled contigs should be approached with caution. In our study, 13,806 assembled gene models (73.07% of the total assembled contigs, Figure 3A) were considered high confidence annotations due to the presence of a significant high stringency BLAST hit in other angiosperm species, as well as a predicted CDS. These contigs had relatively high coverage per base (CPB) values (median 47X) as compared to contigs lacking a predicted CDS (median CPB of 20× or lower, Figure 3B and 3D and Supplemental Table S3). Thus, a lack of CDS prediction was generally associated with low gene expression level and low CPB, which resulted in 'N's inserted by Velvet in the contig sequences (Figure 3B and 3D and Supplemental Table S3). The assembly quality and annotation of these sequences could be improved in future by even deeper sequencing and the addition of data from new tissue types. Another possible source of error is the spurious prediction of CDSs in long, non-coding RNAs, which has been previously shown to occur [44,45]. It is notable that of the 1,813 *Eucalyptus*-derived contigs with no significant BLAST hit to other angiosperms, but containing a predicted CDS (Figure 3C), only 81 contigs had predicted InterProScan domains. Additionally, the median CDS to contig length ratio was 0.33, as compared to 0.62 in the 13,806 high confidence contigs in Figure 3A, which suggests that many of these CDS predictions may be false positives. *De novo* assembled transcriptome datasets lack the ability to distinguish and classify the lower confidence annotations, an exercise that is beyond the scope of this study, albeit one that can be resolved once a genome-based predicted set of gene models is available.

Validation of the digital expression (FPKM) profiles using the "core xylem gene set" identified in *Arabidopsis* [42] has precedence in similar investigations in conifers [46], cotton [47] and poplar [43]. This analysis, combined with the results shown in Figure 5A and Figure

5B, lend support to the biological significance of digital expression profiles derived from short-read sequencing technology, which will assist in the discovery and annotation of novel *Eucalyptus* genes - and using the genome sequence, promoters - playing key roles in growth and development, and particularly in woody biomass production. The Eucspresso online resource produced from this study, as well as future comparative analysis with other woody species such as *Vitis* and *Populus*, will be valuable for studying the unique biology of woody perennials.

## Conclusions

Taking into consideration the number, length, coverage and quality of assembled gene models, as well as their digital expression profiles, this dataset surpasses several previous *de novo* transcriptome assemblies using Illumina [17,18] or 454 technology [13-16]. This can primarily be attributed to the amount of data generated (3.93 Gbp of non-rRNA derived reads), the diversity of tissues sampled and strategy of paired-end sequencing, as well as read-length (mostly 50-60 bp, compared to only 36 bp in earlier studies). Our dataset was generated using several generations of Illumina GA technology, but considering the current throughput of Illumina sequencing (up to 100 Gbp per flowcell), a gene catalog of this scale can now be produced using a single lane of Illumina mRNA-Seq. Finally, non-normalized short-read data will be extremely useful for downstream applications such as digital gene expression profiling and detection of alternative transcript structure, once reference models are available from the genome.

## Methods

### Plant tissue collection

Tissues from a six-year-old ramet of a commercially grown *E. grandis* × *E. urophylla* hybrid clone (GUSAP1, Sappi Forestry, Kwambonambi, South Africa) were collected in a clonal field trial and immediately frozen in liquid nitrogen, as previously described by Ranik and Myburg [48]. The following tissues were sampled from approximately breast height (1.35 m) on the main stem following bark removal: immature xylem (outer glutinous 1-2 mm layer comprising early developing xylem tissue) and xylem (after removal of the immature xylem layer, 2-mm-deep planing including xylem cells in advanced stages of maturity). Early developing phloem tissue including small amounts of cambial cells was collected by scraping the first 1-2 mm layer from the inner surface of the bark. Additionally, we sampled shoot tips (soft green termini of young crown tip branches containing shoot primordia and apical meristems), young leaves (rapidly-growing leaves in the process of unfolding) and mature leaves (older, fully expanded leaves of the current growth season).

### Paired-end mRNA-Seq library preparation and sequence generation

Total RNA was extracted from the six tissues using the protocol described previously [49]. Total RNA quality and concentration were determined using the Agilent RNA 6000 Pico kit (Agilent, Santa Clara, CA) on a 2100 Bioanalyzer (Agilent). Enrichment of polyA+ RNA was performed using the Oligotex midi kit (Qiagen, Valencia, CA). Two hundred nanograms of polyA+ RNA were fragmented in 1× RNA fragmentation solution (Ambion, Austin, TX) at 70°C for 5 minutes. The fragmented RNA was precipitated with three volumes of ethanol and re-dissolved in water. Double-stranded cDNA was synthesized using the cDNA Synthesis System (Roche, Indianapolis, IN) according to manufacturer's instructions using random hexamers (Invitrogen, Carlsbad, CA) to prime the first strand cDNA synthesis. Paired-end libraries with approximate average insert lengths of 200 base pairs were synthesized using the Genomic Sample Prep kit (Illumina, San Diego, CA) according to manufacturer's instructions. Prior to cluster generation, library concentration and size were assayed using the Agilent DNA1000 kit (Agilent) on a 2100 Bioanalyzer (Agilent). Libraries were sequenced on a Genome Analyzer equipped with a paired-end module (versions I, II and IIx, Illumina).

### *De novo* assembly of mRNA-Seq data

After removing sequences containing low quality bases ('N's) or single base repeats and ribosomal RNA sequences, the 3.93 Gbp dataset was used for assembly and subsequent coverage per base (CPB) estimation for each assembled contig. We assembled the filtered Illumina paired-end (PE) reads using Velvet version 0.7.30 [29]. Previous studies [1-3,50] have demonstrated that mRNA-Seq technology produces uneven coverage over a transcript, which prompted us to follow a coverage-assisted reference assembly strategy. Using Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) to align the filtered Illumina PE sequences to the assembled contigs, the average coverage per contig was calculated. A custom script was then developed to extract the pairs of sequences that mapped to each contig, and using that contig as a template, each contig was re-assembled using Velvet with the associated expected coverage parameter set to the Mosaik average coverage value for that contig.

### Contig validation

The degree to which the assembled contigs represented long, contiguous RNA transcript sequences, was evaluated by aligning 35 Velvet contigs and their respective predicted CDSs to full-length, cloned, Sanger-derived *Eucalyptus* reference sequences present in NCBI. CPB was calculated for the sequences using BWA [51] and a global pairwise alignment of the sequences was performed using the Needle package from EMBOSS [52]. Plots were constructed from the alignments with the CPB on the y-axis of the plot. Zero coverage values were assigned to gaps in the alignments. This revealed where gaps and/or potentially misassembled regions were present in the assembled contigs, and to what depth these contigs were sequenced.

### Coding sequence prediction

Coding sequence predictions were performed using GENSCAN [30] and AUGUSTUS [53], predicting 15,713 and 15,904 proteins respectively. The difference in coding sequences predicted could be attributed to the different training data sets used and inherent difficulty of predicting coding sequences from incomplete genomic sequences. The GENSCAN results (15,713 predicted proteins) were used in downstream analyses.

### Annotation of assembled contigs

Homology searches were performed against public sequence databases. The newest versions as of February 2010 of the protein sequences of *Arabidopsis* (TAIR 9), *Vitis* (Sept 2009 build) and *Populus* (version 2.0, Phytozome) were used to construct the individual BLAST datasets. The *Eucalyptus* public dataset (EucAll) consisted of 45,442 entries in Genbank (downloaded March 2010), 13,930 entries from the *Eucalyptus* Wood unigenes and ESTs [33], *E. grandis* leaf tissue ESTs (120,661 entries from DOE-JGI-produced 454 sequences, http://eucalyptusdb.bi.up.ac.za/) and 190,106 Unigenes and singlets from *E. grandis* 454 data [15]. The BLAST e-value threshold was set at $1e^{-10}$, with a minimum alignment length of 100 nucleotides (33 amino acids). Functional annotation (GO and KEGG) was performed using BLAST2GO [54], using the default annotation parameters (BLAST e-value threshold of $1e^{-06}$, Gene Ontology annotation threshold of 55). InterPro annotations were performed using InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/).

### Coverage and FPKM determination

Sequence depth and base coverage were calculated using BWA (Lin et al. 2009) and the FPKM values estimated by aligning the Illumina reads to the assembled transcriptome using Bowtie [40] and estimating the expression level of each predicted transcript (FPKM value) using Cufflinks (http://cufflinks.cbcb.umd.edu) [41].

## Additional material

**Additional file 1: Supplemental Tables S1-S3 and Supplemental Figures S1-S8 referred to in text.**

**Additional file 2: FASTA formatted sequences of all 18,894 assembled contigs**.

**Additional file 3: Contig validation, Needleman-Wunsch alignment figures**.

**Additional file 4: Table containing all 18,894 contig names and calculated FPKM values for six tissues (immature xylem, xylem, phloem, shoot-tips, young leaves and mature leaves)**. Eucspresso (http://eucspresso.bi.up.ac.za/) - Online database with mRNA contig sequences and their Blast, GO, KEGG, Pfam annotations. The short-read sequence data have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession SRA012408.

**Acknowledgements**
The authors would like to acknowledge J. Rees and J.-M. Celton of the University of the Western Cape (Cape Town, South Africa) for assistance with Illumina GA sequencing. Plant materials were kindly provided by Sappi Forestry (Kwambonambi, South Africa). This work was supported through a strategic research grant from the South African Department of Science and Technology (DST) and by research funding from Sappi and Mondi, through the Wood and Fibre Molecular Genetics (WFMG) Programme, the Technology and Human Resources for Industry Programme (THRIP) and the National Research Foundation (NRF) of South Africa.

**Author details**
[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa. [2]Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Pretoria, 0002, South Africa.

### Authors' contributions

EM drafted the manuscript, helped sample the material, prepared the libraries, participated in the *de novo* assembly and data analysis, and helped design Eucspresso. CAH performed the *de novo* assembly and automated annotation, participated in data analysis, designed the database Eucspresso, and helped draft the manuscript. MR prepared the libraries, helped sample the material and participated in data analysis. FJ participated in data analysis and the design of Eucspresso. AAM conceived of the study, and participated in its design and coordination and helped to draft the manuscript and participated in data analysis, and helped design Eucspresso. It is the authors' opinion than EM and CAH contributed equally as first authors to this manuscript. All authors have read and approved the final version of the manuscript.

Received: 30 May 2010 Accepted: 1 December 2010
Published: 1 December 2010

### References

1. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, *et al*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5(7)**:613-619.
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5(7)**:621-628.
3. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320(5881)**:1344-1349.
4. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, *et al*: **mRNA-Seq whole-transcriptome analysis of a single cell.** *Nat Methods* 2009, **6(5)**:377-382.
5. Wilhelm BT, Landry JR: **RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing.** *Methods* 2009, **48(3)**:249-257.
6. Farrer RA, Kemen E, Jones JDG, Studholme DJ: **De novo assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads: RESEARCH LETTER.** *FEMS Microbiol Lett* 2009, **291(1)**:103-111.

7. Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J: ***De novo* bacterial genome sequencing: Millions of very short reads assembled on a desktop computer.** *Genome Res* 2008, **18(5)**:802-809.
8. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ: **Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.** *Nat Methods* 2009, **6(4)**:291-295.
9. DiGuistini S, Liao N, Platt D, Robertson G, Seidel M, Chan S, Docking TR, Birol I, Holt R, Hirst M: ***De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data.** *Genome Biology* 2009, **10(9)**.
10. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC: **De novo Assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis.** *PLoS Genet* 2010, , **6**: e1000891.
11. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, *et al*: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2010, **463(7279)**:311-317.
12. Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM: **Shedding light on an extremophile lifestyle through transcriptomics.** *New Phytol* 2009, **183(3)**:764-775.
13. Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL: **Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*.** *BMC Genomics* 2009, **10(234)**.
14. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **10(219)**.
15. Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome.** *BMC Genomics* 2008, **9(312)**.
16. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17(7)**:1636-1647.
17. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, *et al*: ***De novo* transcriptome assembly with ABySS.** *Bioinformatics* 2009, **25(21)**:2872-2877.
18. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A: **Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics.** *Mol Biol Evol* 2009, **26(12)**:2731-2744.
19. Wu T, Qin Z, Zhou X, Feng Z, Du Y: **Transcriptome profile analysis of floral sex determination in cucumber.** *J Plant Physiol* 2010, **167(11)**:905-913.
20. Eldridge K, Davidson J, Harwood C, van Wyk G: *Eucalypt domestication and breeding* Oxford: Clarendon Press; 1993.
21. FAO: **Forests and Energy.** *FAO Forestry Paper No* 2008, **154, (Rome):(ISBN 978-992-975-105985-105982).**
22. Hinchee M, Rottmann W, Mullinax L, Zhang C, Chang S, Cunningham M, Pearson L, Nehra N: **Short-rotation woody crops for bioenergy and biofuels applications.** *In Vitro Cell Dev Biol - Plant* 2009, **45(6)**:619-629.
23. Myburg AA, Grattapaglia D, Tuskan GA, Schmutz J, Barry K, Bristow J, The Eucalyptus Genome Network: **Sequencing the *Eucalyptus* genome: Genomic resources for renewable energy and fiber production.** *Plant & Animal Genome XVI Conference: January 12-16, 2008; San Diego, CA* 2008.
24. Hibino T: **"Post-genomics" research in *Eucalyptus* in the near future.** *Plant Biotechnol* 2009, **26(1)**:109-113.
25. Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin X, *et al*: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408(6814)**:796-815.
26. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam M, Ralph S, Rombauts S, Salamov A, *et al*: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313(5793)**:1596-1604.
27. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449(7161)**:463-467.
28. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S: **The *Populus* Genome Integrative Explorer (PopGenIE): A new resource for exploring the *Populus* genome.** *New Phytol* 2009, **182(4)**:1013-1025.
29. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18(5)**:821-829.
30. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268(1)**:78-94.
31. Pinheiro M, Afreixo V, Moura G, Freitas A, Santos MAS, Oliveira JL: **Statistical, computational and visualization methodologies to unveil gene primary structure features.** *Methods Inf Med* 2006, **45(2)**:163-168.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3)**:403-410.
33. Rengel D, Clemente HS, Servant F, Ladouce N, Paux E, Wincker P, Couloux A, Sivadon P, Grima-Pettenati J: **A new genomic resource dedicated to wood formation in *Eucalyptus*.** *BMC Plant Biol* 2009, **9(36)**.
34. Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, *et al*: **Poplar carbohydrate-active enzymes. Gene identification and expression analyses.** *Plant Physiol* 2006, **140(3)**:946-962.
35. Pavy N, Laroche J, Bousquet J, Mackay J: **Large-scale statistical analysis of secondary xylem ESTs in pine.** *Plant Mol Biol* 2005, **57(2)**:203-224.
36. Hale MC, McCormick CR, Jackson JR, DeWoody JA: **Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): The relative merits of normalization and rarefaction in gene discovery.** *BMC Genomics* 2009, **10(203)**.
37. Kristiansson E, Asker N, Förlin L, Joakim DGJ: **Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing.** *BMC Genomics* 2009, **10(345)**.
38. Schwarz D, Robertson HM, Feder JL, Varala K, Hudson ME, Ragland GJ, Hahn DA, Berlocher SH: **Sympatric ecological speciation meets pyrosequencing: Sampling the transcriptome of the apple maggot *Rhagoletis pomonella*.** *BMC Genomics* 2009, **10(633)**.
39. Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB: **Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing.** *Plant Physiol* 2007, **144(1)**:32-42.
40. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009, **10(3)**.
41. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28(5)**:511-515.
42. Ko JH, Beers EP, Han KH: **Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*.** *Mol Genet Genomics* 2006, **276(6)**:517-531.
43. Dharmawardhana P, Brunner AM, Strauss SH: **Genome-wide transcriptome analysis of the transition from primary to secondary stem development in *Populus trichocarpa*.** *BMC Genomics* 2010, **11(1)**:150.
44. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES: **Distinguishing protein-coding and noncoding genes in the human genome.** *Proc Natl Acad Sci USA* 2007, **104(49)**:19428-19433.
45. Dinger ME, Pang KC, Mercer TR, Mattick JS: **Differentiating protein-coding and noncoding RNA: Challenges and ambiguities.** *PLoS Comput Biol* 2008, **4(11)**:1-5.
46. Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H, *et al*: **Identification of conserved core xylem gene sets: Conifer cDNA microarray development, transcript profiling and computational analyses.** *New Phytol* 2008, **180(4)**:766-786.
47. Betancur L, Singh B, Rapp RA, Wendel JF, Marks MD, Roberts AW, Haigler CH: **Phylogenetically distinct cellulose synthase genes support secondary wall thickening in arabidopsis shoot trichomes and cotton fiber.** *J Integr Plant Biol* 2010, **52(2)**:205-220.
48. Ranik M, Myburg AA: **Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis.** *Tree Physiol* 2006, **26(5)**:545-556.
49. Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees.** *Plant Mol Biol Report* 1993, **11(2)**:113-116.
50. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*.** *Cell* 2008, **133(3)**:523-536.
51. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25(14)**:1754.
52. Rice P, Longden I, Bleasby A: **EMBOSS: the European molecular biology open software suite.** *Trends Genet* 2000, **16(6)**:276-277.
53. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19(SUPPL 2)**:ii215-ii225.

54. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21(18)**:3674-3676.

doi:10.1186/1471-2164-11-681
**Cite this article as:** Mizrachi *et al.*: *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* 2010 **11**:681.

# Bibliography

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B. and Moreno, R. F. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project *Science* **252**, 5013, 1651–6. 3, 82

AGI, The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana *Nature* **408**, 6814, 796–815. 16

Ahn, S.-M., Kim, T.-H., Lee, S., Kim, D., Ghang, H., Kim, D.-S., Kim, B.-C., Kim, S.-Y., Kim, W.-Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S., Kim, C. G., Cha, J.-Y., Kim, K.-H., Lee, B., Bhak, J. and Kim, S.-J. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**, 9, 1622–1629 ISSN 1549-5469 (Electronic); 1088-9051 (Linking). 18

Akhunov, E., Nicolet, C. and Dvorak, J. (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay *Theor Appl Genet* **119**, 3, 507–17. 18

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 3, 403–410. 30, 46, 48, 87

Andersson-Gunnerås, S., Mellerowicz, E. J., Love, J., Segerman, B., Ohmiya, Y., Coutinho, P. M., Nilsson, P., Henrissat, B., Moritz, T. and Sundberg, B. (2006) Biosynthesis of cellulose-enriched tension wood in Populus: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis *Plant J* **45**, 2, 144–65. 108

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P. and Nordborg,

M. (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes *PLoS Genet* **1**, 5, e60. 18

Arnaiz, O., Goût, J.-F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E. and Sperling, L. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia *BMC Genomics* **11**, 547. 79

Aury, J.-M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F. and Wincker, P. (2008) High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies *BMC Genomics* **9**, 603. 15

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. and Johnson, E. A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, 10, e3376. 19

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L. and Schnable, P. S. (2007) SNP discovery via 454 transcriptome sequencing *Plant J* **51**, 5, 910–8. 19

Barbazuk, W. B., Fu, Y. and McGinnis, K. M. (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges *Genome Res* **18**, 9, 1381–92. 23

Batzoglou, S. (2005) *Encyclopedia of genomics, proteomics and bioinformatics* chapter Algorithmic Challenges in Mammalian Genome Sequence Assembly John Wiley and Sons. 28

Bayer, E. M., Bottrill, A. R., Walshaw, J., Vigouroux, M., Naldrett, M. J., Thomas, C. L. and Maule, A. J. (2006) Arabidopsis cell wall proteome defined using multidimensional protein identification technology *Proteomics* **6**, 1, 301–11. 116

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance,

P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 7218, 53–59 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18

Bertone, P., Gerstein, M. and Snyder, M. (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res* **13**, 3, 259–274 ISSN 0967-3849

(Print); 0967-3849 (Linking). 20

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. and Snyder, M. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 5705, 2242–2246 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 20, 21

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., Horsman, D. E., Connors, J. M., Gascoyne, R. D., Marra, M. A. and Jones, S. J. M. (2009) De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 21, 2872–2877 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 29, 118, 142

Bischoff, V., Nita, S., Neumetzler, L., Schindelasch, D., Urbain, A., Eshed, R., Persson, S., Delmer, D. and Scheible, W.-R. (2010) TRICHOME BIREFRINGENCE and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in Arabidopsis *Plant Physiol* **153**, 2, 590–602. 116

Blanca, J., Cañizares, J., Roig, C., Ziarsolo, P., Nuez, F. and Picó, B. (2011) Transcriptome characterization and high throughput SSRs and SNPs discovery in Cucurbita pepo (Cucurbitaceae) *BMC Genomics* **12**, 104. 79

Blanco, E. and Guigó, R. (2005) *Bioinformatics: A practical guide to the analysis of genes and proteins* chapter Predictive methods using DNA sequences, 116–142 3 John Wiley and Sons. 119

Bloom, J. S., Khan, Z., Kruglyak, L., Singh, M. and Caudy, A. A. (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays *BMC Genomics* **10**, 221. 80

Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993) dbEST–database for "expressed sequence tags" *Nat Genet* **4**, 4, 332–3. 82

Boguski, M. S., Tolstoshev, C. M. and Bassett, D. E., Jr (1994) Gene discovery in dbEST *Science* **265**, 5181, 1993–4. 82

Bohnert, R. and Rätsch, G. (2010) rQuant.web: a tool for RNA-Seq-based transcript quantitation *Nucleic*

*Acids Res* **38**, Web Server issue, W348–51. 80

Bokhari, S. H. and Sauer, J. R. (2005) A parallel graph decomposition algorithm for DNA sequencing with nanopores *Bioinformatics* **21**, 7, 889–96. 28

Borodovsky, M. and McIninch, J. (1993) GENMARK: parallel gene recognition for both DNA strands *Comput Chem* **17**, 2, 123–133. 86

Bosca, S., Barton, C. J., Taylor, N. G., Ryden, P., Neumetzler, L., Pauly, M., Roberts, K. and Seifert, G. J. (2006) Interactions between MUR10/CesA7-dependent secondary cellulose biosynthesis and primary cell wall structure *Plant Physiol* **142**, 4, 1353–63. 116

Bowers, J., Mitchell, J., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G. M., Marappan, S., McInerney, P., Platt, A., Roy, A., Siddiqi, S. M., Steinmann, K. and Thompson, J. F. (2009) Virtual terminator nucleotides for next-generation DNA sequencing. *Nat Methods* **6**, 8, 593–595 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 13

Boyle, J. (2004) Bioinformatics in undergraduate education: Practical examples *Biochemistry and Molecular Biology Education* **32**, 4, 236–238. 43

Braslavsky, I., Hebert, B., Kartalov, E. and Quake, S. R. (2003) Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* **100**, 7, 3960–3964 ISSN 0027-8424 (Print); 0027-8424 (Linking). 4, 12

Brem, R. B. and Kruglyak, L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast *Proc Natl Acad Sci U S A* **102**, 5, 1572–7. 145

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* **18**, 6, 630–634 ISSN 1087-0156 (Print); 1087-0156 (Linking). 20

Brown, D. M., Zeef, L. A. H., Ellis, J., Goodacre, R. and Turner, S. R. (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics *Plant Cell* **17**, 8, 2281–95. 108, 116

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 1, 78–94 ISSN 0022-2836 (Print); 0022-2836 (Linking). 46, 71, 86

Burrows, M. and Wheeler, D. (1994) A block-sorting lossless data compression algorithm Technical Report 124 Digital Equipment Corporation. 32

Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M. and Buell, C. R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis *BMC Genomics* **7**, 327. 23

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A. and Hayashizaki, Y. (2006) Genome-wide analysis of mammalian promoter architecture and evolution *Nat Genet* **38**, 6, 626–35. 21

Cartieaux, F., Thibaud, M.-C., Zimmerli, L., Lessard, P., Sarrobert, C., David, P., Gerbaud, A., Robaglia, C., Somerville, S. and Nussaume, L. (2003) Transcriptome analysis of Arabidopsis colonized by a plant-growth promoting rhizobacterium reveals a general effect on disease resistance *Plant J* **36**, 2, 177–88. 108

Casneuf, T., Van de Peer, Y. and Huber, W. (2007) In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics* **8**, 461 ISSN 1471-2105 (Electronic); 1471-2105 (Linking). 20

Chapman, B. and Chang, J. (2000) Biopython: Python tools for computational biology *SIGBIO Newsl* **20**, 2, 15–19 ISSN 0163-5697. 43

Che, P., Lall, S., Nettleton, D. and Howell, S. H. (2006) Gene expression programs during shoot, root, and callus development in Arabidopsis tissue culture *Plant Physiol* **141**, 2, 620–37. 108

Chen, F.-C., Wang, S.-S., Chaw, S.-M., Huang, Y.-T. and Chuang, T.-J. (2007) Plant Gene and Alternatively Spliced Variant Annotator. A plant genome annotation pipeline for rice gene and alternatively spliced variant identification with cross-species expressed sequence tag conservation from seven plant species *Plant Physiol* **143**, 3, 1086–95. 23

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S. and Gingeras, T. R. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 5725, 1149–1154 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 20

Clark, T., Sugnet, C. and Ares, M. (2002) Genomewide analysis of mRNA processing in Yeast using splicing-specific microarrays *Science* **296**, 5569, 907–910. 20

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J. and Grimmond, S. M. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**, 7, 613–619 ISSN 1548-7105 (Electronic). 3, 20, 22, 82, 141

Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., Druka, A., Stein, N., Svensson, J. T., Wanamaker, S., Bozdag, S., Roose, M. L., Moscou, M. J., Chao, S., Varshney, R. K., Szucs, P., Sato, K., Hayes, P. M., Matthews, D. E., Kleinhofs, A., Muehlbauer, G. J., DeYoung, J., Marshall, D. F., Madishetty, K., Fenton, R. D., Condamine, P., Graner, A. and Waugh, R. (2009) Development and implementation of high-throughput SNP genotyping in barley *BMC Genomics* **10**, 582. 18

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 6, 1767–1771 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 26

Coetzer, N., Gazendam, I., Oelofse, D. and Berger, D. K. (2010) SSHscreen and SSHdb, generic software for microarray based gene discovery: application to the stress response in cowpea *Plant Methods* **6**, 10. 79

Cohen, J. (2003) Guidelines for Establishing Undergraduate Bioinformatics Courses *Journal of Science Education and Technology* **12**, 4, 449–456. 43

Collins, L. J., Biggs, P. J., Voelckel, C. and Joly, S. (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences *Genome Inform* **21**, 3–14. 83

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research *Bioinformatics* **21**, 18, 3674–6. 46, 48, 70, 143

Darzentas, N. (2010) Circoletto: visualizing sequence similarity with Circos *Bioinformatics*. 46

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. and Steinmetz, L. M. (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**, 14, 5320–5325 ISSN 0027-8424 (Print); 0027-8424 (Linking). 20

De Bona, F., Ossowski, S., Schneeberger, K. and Ratsch, G. (2008) Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**, 16, i174–80 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 34

De la Vega, F. M., Lazaruk, K. D., Rhodes, M. D. and Wenz, M. H. (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System *Mutat Res* **573**, 1-2, 111–35. 18

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O. and Artiguenave, F. (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biol* **9**, 12, R175 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 3, 20, 22, 34, 82, 141

Dias Neto, E., Correa, R. G., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva, W., Jr, Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., Carvalho, A. F., Matsukuma, A., Baia, G. S., Simpson, D. H., Brunstein, A., de Oliveira, P. S., Bucher, P., Jongeneel, C. V., O'Hare, M. J., Soares, F., Brentani, R. R., Reis, L. F., de Souza, S. J. and Simpson, A. J. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags *Proc Natl Acad Sci U S A* **97**, 7, 3491–6. 82

DiGuistini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan, S., Docking, T., Birol, I., Holt, R., Hirst, M., Mardis, E., Marra, M. A., Hameling, R. C., Bohlmann, J., Breuil, C. and Jones, S. J. M. (2010) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data *Genome Biology* **9**, R94. 15, 141

Dolezel, J., Kubaláková, M., Paux, E., Bartos, J. and Feuillet, C. (2007) Chromosome-based genomics in the cereals *Chromosome Res* **15**, 1, 51–66. 16

Doukhanina, E. V., Chen, S., van der Zalm, E., Godzik, A., Reed, J. and Dickman, M. B. (2006) Identification and functional characterization of the BAG protein family in Arabidopsis thaliana *J Biol Chem* **281**, 27, 18793–801. 108

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. and Vogelstein, B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations *Proc Natl Acad Sci U S A* **100**, 15, 8817–22. 8

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M. and Reid, C. A. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 5961, 78–81 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 4, 9, 10, 18, 141

Durham, A. M., Kashiwabara, A. Y., Matsunaga, F. T. G., Ahagon, P. H., Rainone, F., Varuzza, L. and Gruber, A. (2005) EGene: a configurable pipeline generation system for automated sequence analysis

*Bioinformatics* **21**, 12, 2812–3. 42

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., Lushbough, C. and Brendel, V. (2008) PlantGDB: a resource for comparative plant genomics *Nucleic Acids Res* **36**, Database issue, D959–65. 123

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res* **32**, 5, 1792–7. 46

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 5910, 133–138 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 4, 10, 11

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005) The Sequence Ontology: a tool for the unification of genome annotations *Genome Biol* **6**, 5, R44. 38

Eker, J., Janneck, J., Lee, E. A., Liu, J., Liu, X., Ludvig, J., Sachs, S. and Xiong, Y. (2003) Taming heterogeneity - the Ptolemy approach *Proceedings of the IEEE* **91**, 1, 127–144. 37

Eklund, A. C., Turner, L. R., Chen, P., Jensen, R. V., deFeo, G., Kopf-Sill, A. R. and Szallasi, Z. (2006) Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat Biotechnol* **24**, 9, 1071–1073 ISSN 1087-0156 (Print); 1087-0156 (Linking). 20

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 3, 186–194 ISSN 1088-9051 (Print); 1088-9051 (Linking). 26

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 3, 175–185 ISSN 1088-9051 (Print); 1088-9051 (Linking). 26

Fedurco, M., Romieu, A., Williams, S., Lawrence, I. and Turcatti, G. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* **34**, 3, e22 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 4, 6

Ferragina, P. and Manzini, G. (2000) Opportunistic data structures with applications in *FOCS 2000: Proceedings of the 41st Annual Symposium on Foundations of Computer Science* p390 IEEE Computer Society, Washington, DC, USA. 32

Ferragina, P. and Manzini, G. (2001) An experimental study of an opportunistic index in *SODA 2001: Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms* p269–279 Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. 32

Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., Wong, W.-K. and Mockler, T. C. (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res* **20**, 1, 45–58 ISSN 1549-5469 (Electronic); 1088-9051 (Linking). 23, 83

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. and Bateman, A. (2010) The Pfam protein families database *Nucleic Acids Res* **38**, Database issue, D211–22. 48

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J. and Turner, S. W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 6, 461–465 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 12, 141

Forment, J., Gilabert, F., Robles, A., Conejero, V., Nuez, F. and Blanca, J. M. (2008) EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration *BMC Bioinformatics* **9**, 5. 42

Forrest, M. and Moore, T. (2008) Eucalyptus gunnii: A possible source of bioenergy? *Biomass and Bioenergy* **32**, 10, 978–980. 1

Frey, B. J., Mohammad, N., Morris, Q. D., Zhang, W., Robinson, M. D., Mnaimneh, S., Chang, R., Pan, Q., Sat, E., Rossant, J., Bruneau, B. G., Aubin, J. E., Blencowe, B. J. and Hughes, T. R. (2005) Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. *Nat Genet* **37**, 9, 991–996 ISSN 1061-4036 (Print); 1061-4036 (Linking). 20

Fullwood, M. J., Wei, C.-L., Liu, E. T. and Ruan, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* **19**, 4, 521–532 ISSN 1088-9051 (Print); 1088-9051 (Linking). 14, 19

García-Alcalde, F., García-Lopez, F., Dopazo, J. and Conesa, A. (2010) Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data *Bioinformatics*. 89

Garcia-Hernandez, M., Berardini, T. Z., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Rhee, S. Y., Scholl, R., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2002) TAIR: a resource for integrated Arabidopsis data *Funct Integr Genomics* **2**, 6, 239–53. 123

Garg, R., Patel, R. K., Tyagi, A. K. and Jain, M. (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification *DNA Res* **18**, 1, 53–63. 144

Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., Panhuis, T. M., Mieczkowski, P., Secchi, A., Bosco, D., Berney, T., Montanya, E., Mohlke, K. L., Lieb, J. D. and Ferrer, J. (2010) A map of open chromatin in human pancreatic islets *Nat Genet* **42**, 3, 255–9. 38

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation *Genome Res* **11**, 8, 1425–33. 38, 48

Ghadessy, F. J., Ong, J. L. and Holliger, P. (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A* **98**, 8, 4552–4557 ISSN 0027-8424 (Print); 0027-8424 (Linking). 5

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J. and Nekrutenko, A. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 10, 1451–1455 ISSN 1088-9051 (Print); 1088-9051 (Linking). 44

Gilad, Y., Pritchard, J. K. and Thornton, K. (2009) Characterizing natural variation using next-generation sequencing technologies *Trends Genet* **25**, 10, 463–71. 145

Goecks, J., Nekrutenko, A., Taylor, J. and Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences *Genome Biol* **11**, 8, R86. 35, 38, 142

Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.-l., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica) *Science* **296**, 5565, 92–100. 16

Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P. M. and Bernstein, B. E. (2010) Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat Methods* **7**, 1, 47–49 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 13, 141

Goto, N., Nakao, M., Kawashima, S., Katayama, T. and Kanehisa, M. (2003) BioRuby: open-source bioinformatics library *GENOME INFORMATICS SERIES* 629–630. 43

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome *Nat Biotechnol.* 79, 118, 142

Grattapaglia, D. and Kirst, M. (2008) Eucalyptus applied genomics: from gene sequences to breeding tools *New Phytol* **179**, 4, 911–29. 1

Graveley, B. R. (2008) Molecular biology: Power sequencing *Nature* **453**, 7199, 1197–1198. 3

Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1992) Prediction of gene structure *Journal of Molecular Biology* **226**, 1, 141–157. 87

Hansen, K. D., Brenner, S. E. and Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming *Nucleic Acids Res* **38**, 12, e131. 80

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H. and Xie, Z. (2008) Single-molecule DNA sequencing of a viral genome. *Science* **320**, 5872, 106–109 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 13

Hashimoto, S., Qu, W., Ahsan, B., Ogoshi, K., Sasaki, A., Nakatani, Y., Lee, Y., Ogawa, M., Ametani, A., Suzuki, Y., Sugano, S., Lee, C. C., Nutter, R. C., Morishita, S. and Matsushima, K. (2009) High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. *PLoS ONE* **4**, 1, e4108. 141

Hibino, T. (2009) "Post-genomics" research in Eucalyptus in the near future *Plant Biotechnology* **26**, 1, 109–113. 82

Hiller, D., Jiang, H., Xu, W. and Wong, W. H. (2009) Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics* **25**, 23, 3056–3059 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 21, 79

Hinchee, M., Rottmann, W., Mullinax, L., Zhang, C., Chang, S., Cunningham, M., Pearson, L. and Nehra, N. (2009) Short-rotation woody crops for bioenergy and biofuels applications. *In Vitro Cell Dev Biol Plant* **45**, 6, 619–629 ISSN 1054-5476 (Print); 1054-5476 (Linking). 2, 82

Hofreuter, D., Tsai, J., Watson, R. O., Novik, V., Altman, B., Benitez, M., Clark, C., Perbost, C., Jarvie, T., Du, L. and Galan, J. E. (2006) Unique features of a highly pathogenic Campylobacter jejuni strain. *Infect Immun* **74**, 8, 4694–4707 ISSN 0019-9567 (Print); 0019-9567 (Linking). 15, 16, 141

Holland, R. C. G., Down, T. A., Pocock, M., Prlić, A., Huen, D., James, K., Foisy, S., Dräger, A., Yates, A., Heuer, M. and Schreiber, M. J. (2008) BioJava: an open-source framework for bioinformatics *Bioinformatics* **24**, 18, 2096–7. 43

Homer, N., Merriman, B. and Nelson, S. F. (2009*a*) BFAST: an alignment tool for large scale genome resequencing *PLoS One* **4**, 11, e7767. 31

Homer, N., Merriman, B. and Nelson, S. F. (2009*b*) Local alignment of two-base encoded DNA sequence *BMC Bioinformatics* **10**, 175. 31

Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L. A., Bhattacharyya, D., Bhaya, D., Sobral, B. W., Beavis, W., Meinke, D. W., Town, C. D., Somerville, C. and Rhee, S. Y. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* **29**, 1, 102–105 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 88

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W. J., Wang, X., Xie, B., Ni, P., Ren, Y., Zhu, H., Li, J., Lin, K., Jin, W., Fei, Z., Li, G., Staub, J., Kilian, A., van der Vossen, E. A. G., Wu, Y., Guo, J., He, J., Jia, Z., Ren, Y., Tian, G., Lu, Y., Ruan, J., Qian, W., Wang, M., Huang, Q., Li, B., Xuan, Z., Cao, J., Asan, Wu, Z., Zhang, J., Cai, Q., Bai, Y., Zhao, B., Han, Y., Li, Y., Li, X., Wang, S., Shi, Q., Liu, S., Cho, W. K., Kim, J.-Y., Xu, Y., Heller-Uszynska, K., Miao, H., Cheng, Z., Zhang, S., Wu, J., Yang, Y., Kang, H., Li, M., Liang, H., Ren, X., Shi, Z., Wen, M., Jian, M., Yang, H., Zhang, G., Yang, Z., Chen, R., Liu, S., Li, J., Ma, L., Liu, H., Zhou, Y., Zhao, J., Fang, X., Li, G., Fang, L., Li, Y., Liu, D., Zheng, H., Zhang, Y., Qin, N., Li, Z., Yang, G., Yang, S., Bolund, L., Kristiansen, K., Zheng, H., Li, S., Zhang, X., Yang, H., Wang, J., Sun, R., Zhang, B., Jiang, S., Wang, J., Du, Y. and Li, S. (2009*a*) The genome of the cucumber, Cucumis sativus L *Nat Genet* **41**, 12, 1275–81. 16

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T. and Han, B. (2009*b*) High-throughput genotyping by whole-genome resequencing. *Genome Res* **19**, 6, 1068–1076 ISSN 1088-9051 (Print); 1088-9051 (Linking). 17

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H. and Yeats, C. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, Database issue, D211–5 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 88

Hyten, D. L., Cannon, S. B., Song, Q., Weeks, N., Fickus, E. W., Shoemaker, R. C., Specht, J. E., Farmer, A. D., May, G. D. and Cregan, P. B. (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11**, 38 ISSN 1471-2164 (Electronic); 1471-2164 (Linking). 17

Hyten, D. L., Song, Q., Choi, I.-Y., Yoon, M.-S., Specht, J. E., Matukumalli, L. K., Nelson, R. L., Shoemaker, R. C., Young, N. D. and Cregan, P. B. (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean *Theor Appl Genet* **116**, 7, 945–52. 18

Illumina (2008) Sequence analysis Software User Guide: For Pipeline Version 1.3 and CASAVA Version 1.0 Technical report Illumina, Inc. 26

Imelfort, M. and Edwards, D. (2009) De novo sequencing of plant genomes using second-generation technologies *Briefings in bioinformatics* **10**, 6, 609. 16

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.-F., Weissenbach, J., Quetier, F. and Wincker, P. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 7161, 463–467 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 88

Jiang, Z., Tang, H., Ventura, M., Cardone, M. F., Marques-Bonet, T., She, X., Pevzner, P. A. and

Eichler, E. E. (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution *Nat Genet* **39**, 11, 1361–8. 28

Johnson, J. M., Edwards, S., Shoemaker, D. and Schadt, E. E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments *Trends Genet* **21**, 2, 93–102. 21

Kapur, K., Xing, Y., Ouyang, Z. and Wong, W. H. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biol* **8**, 5, R82 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 20

Katagiri, T., Ishiyama, K., Kato, T., Tabata, S., Kobayashi, M. and Shinozaki, K. (2005) An important role of phosphatidic acid in ABA signaling during germination in Arabidopsis thaliana *Plant J* **43**, 1, 107–17. 108

Kent, W. J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res* **12**, 4, 656–664 ISSN 1088-9051 (Print); 1088-9051 (Linking). 30

Kim, J.-I., Ju, Y. S., Park, H., Kim, S., Lee, S., Yi, J.-H., Mudge, J., Miller, N. A., Hong, D., Bell, C. J., Kim, H.-S., Chung, I.-S., Lee, W.-C., Lee, J.-S., Seo, S.-H., Yun, J.-Y., Woo, H. N., Lee, H., Suh, D., Lee, S., Kim, H.-J., Yavartanoo, M., Kwak, M., Zheng, Y., Lee, M. K., Park, H., Kim, J. Y., Gokcumen, O., Mills, R. E., Zaranek, A. W., Thakuria, J., Wu, X., Kim, R. W., Huntley, J. J., Luo, S., Schroth, G. P., Wu, T. D., Kim, H., Yang, K.-S., Park, W.-Y., Kim, H., Church, G. M., Lee, C., Kingsmore, S. F. and Seo, J.-S. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 7258, 1011–1015 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18, 19

Kislyuk, A., Katz, L., Agrawal, S. and Hagen, M. (2005) A computational genomics pipeline for microbial sequencing projects. 4

Ko, J.-H., Beers, E. P. and Han, K.-H. (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in Arabidopsis thaliana *Mol Genet Genomics* **276**, 6, 517–31. 108

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y. and Carninci, P. (2006) CAGE: cap analysis of gene expression.

*Nat Methods* **3**, 3, 211–222 ISSN 1548-7091 (Print); 1548-7091 (Linking). 20

Kosakovsky Pond, S., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W.-Y., Taylor, J., Nekrutenko, A. and Galaxy Team (2009) Windshield splatter analysis with the Galaxy metagenomic pipeline *Genome Res* **19**, 11, 2144–53. 38

Külheim, C., Yeoh, S. H., Maintz, J., Foley, W. J. and Moran, G. F. (2009) Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways *BMC Genomics* **10**, 452. 120

Kuznetsov, V. A. (2009) Relative avidity, specificity, and sensitivity of transcription factor-DNA binding in genome-scale experiments *Methods Mol Biol* **563**, 15–50. 141

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S. and Chen, Y. J. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 6822, 860–921 ISSN 0028-0836 (Print); 0028-0836 (Linking). 18

Langmead, B., Hansen, K. D. and Leek, J. T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna *Genome Biol* **11**, 8, R83. 80

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, 3, R25 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 30, 31, 32, 33, 89, 90

Lawrence, C. J., Dong, Q., Polacco, M. L., Seigfried, T. E. and Brendel, V. (2004) MaizeGDB, the community database for maize genetics and genomics *Nucleic Acids Res* **32**, Database issue, D393–7.

Lee, E. A. (2009) Finite State Machines and Modal Models in Ptolemy II Technical Report UCB/EECS-2009-151 EECS Department, University of California, Berkeley. 37

Lee, E. A. and Zheng, H. (2005) Operational Semantics of Hybrid Systems in *HSCC* 25–53. 37

Leung, M.-K., 0002, T. M., Lee, E. A., Latronico, E., Shelton, C. P., Tripakis, S. and Lickly, B. (2009) Scalable Semantic Annotation Using Lattice-Based Ontologies in *MoDELS* 393–407. 37

Lewis, B. P., Green, R. E. and Brenner, S. E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100**, 1, 189–192. 23

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D. E., Koboldt, D. C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L. W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J. R., Minx, P., Gordon, D., Chinwalla, A., Zhao, Y., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W. D., Nagarajan, R., Walter, M. J., Link, D. C., Graubert, T. A., DiPersio, J. F. and Wilson, R. K. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 7218, 66–72 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17

Li, D., Guo, Y., Shao, H., Tellier, L. C., Wang, J., Xiang, Z. and Xia, Q. (2010*a*) Genetic diversity, molecular phylogeny and selection evidence of the silkworm mitochondria implicated by complete resequencing of 41 genomes. *BMC Evol Biol* **10**, 81 ISSN 1471-2148 (Electronic); 1471-2148 (Linking). 17

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 14, 1754–1760 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31, 33, 87, 89

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009*a*) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 16, 2078–2079 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 89, 90

Li, H., Lovci, M. T., Kwon, Y.-S., Rosenfeld, M. G., Fu, X.-D. and Yeo, G. W. (2008*a*) Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci U S A* **105**, 51, 20179–20184 ISSN 1091-6490 (Electronic); 0027-8424 (Linking). 21, 79

Li, H., Ruan, J. and Durbin, R. (2008*b*) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 11, 1851–1858 ISSN 1088-9051 (Print); 1088-9051 (Linking). 30, 33, 90

Li, H., Ruan, J. and Durbin, R. (2008*c*) Mapping short DNA sequencing reads and calling variants using mapping quality scores *Genome Res* **18**, 11, 1851–8. 31

Li, R., Fan, W., Tian, G., Zhu, L., H.and He, Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O., Leung, F.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C., Lam, T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J. and Wang, J. (2010*b*) The sequence and de novo assembly of the giant panda genome *Nature* **463**, 7279, 311–317. 15, 141

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008*d*) SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 5, 713–714 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. and Wang, J. (2009*b*) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 15, 1966–1967 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31, 33

Lin, H., Zhang, Z., Zhang, M. Q., Ma, B. and Li, M. (2008) ZOOM! Zillions of oligos mapped. *Bioinformatics* **24**, 21, 2431–2437 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30

Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., Thompson, J. F., Letovsky, S., Milos, P. and Causey, M. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* **27**, 7, 652–658 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 13

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis *Cell* **133**, 3, 523–36. 141

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J. and Zhao, Y. (2005) Scientific Workflow Management and the Kepler System *Concurrency and Computation: Practice and Experience* 1–19. 35, 37, 43, 44

Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks *Bioinformatics* **21**, 16, 3448–9. 89

Malhis, N., Butterfield, Y. S. N., Ester, M. and Jones, S. J. M. (2009) Slider–maximum use of probability information for alignment of short sequence reads and SNP detection *Bioinformatics* **25**, 1, 6–13. 31

Manfield, I. W., Jen, C.-H., Pinney, J. W., Michalopoulos, I., Bradford, J. R., Gilmartin, P. M. and Westhead, D. R. (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis *Nucleic Acids Res* **34**, Web Server issue, W504–9. 139

Mardis, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 3, 133–141 ISSN 0168-9525 (Print). 3

Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D., Fulton, L. A., Locke, D. P., Magrini, V. J., Abbott, R. M., Vickery, T. L., Reed, J. S., Robinson, J. S., Wylie, T., Smith, S. M., Carmichael, L., Eldred, J. M., Harris, C. C., Walker, J., Peck, J. B., Du, F., Dukes, A. F., Sanderson, G. E., Brummett, A. M., Clark, E., McMichael, J. F., Meyer, R. J., Schindler, J. K., Pohl, C. S., Wallis, J. W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M. E., Ivy, J. V., Kalicki, J., Elliott, G., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M. A., Baty, J., Heath, S., Shannon, W. D., Nagarajan, R., Link, D. C., Walter, M. J., Graubert, T. A., DiPersio, J. F., Wilson, R. K. and Ley, T. J. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 11, 1058–1066 ISSN 1533-4406 (Electronic); 0028-4793 (Linking). 17

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 7057, 376–380 ISSN 1476-4687 (Electronic). 2, 4, 5, 6, 15, 16, 141

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 9, 1509–1517 ISSN 1088-9051 (Print); 1088-9051 (Linking). 21, 79, 80, 139

McKernan, K., Blanchard, A., Kotler, L. and Costa, G. (2006) Reagents, methods and libraries for bead-based sequencing. Technical report filed as US patent application 20080003571. 8

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M. and Blanchard, A. P. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 9, 1527–1541 ISSN 1549-5469 (Electronic); 1088-9051 (Linking). 18, 19, 141

Mir, K. U. (2009) Sequencing genomes: from individuals to populations. *Brief Funct Genomic Proteomic* **8**, 5, 367–378 ISSN 1477-4062 (Electronic); 1473-9550 (Linking). 17

Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T. and Goble, C. (2009) Taverna, reloaded. unpublished technical documentation. 37

Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F. and Myburg, A. A. (2010) De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq *BMC Genomics* **11**, 1, 681. 84, 120, 123, 155

Mondego, J. M., Vidal, R. O., Carazzolle, M. F., Tokuda, E. K., Parizzi, L. P., Costa, G. G., Pereira, L. F., Andrade, A. C., Colombo, C. A., Vieira, L. G., Pereira, G. A. and Brazilian Coffee Genome Project Consortium (2011) An EST-based analysis identifies new genes and reveals distinctive gene expression features of Coffea arabica and Coffea canephora *BMC Plant Biol* **11**, 30. 79

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E. T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 7289, 773–777 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 24

Moore, B., Fan, G. and Eilbeck, K. (2010) SOBA: sequence ontology bioinformatics analysis *Nucleic Acids Res* **38 Suppl**, W161–4. 82

Moore, P., Ming, R. and Delmer, D. (2008) *Genomics of tropical crop plants* volume 1 of *Plant genetics and genomics: Crops and models* chapter Genomics of Eucalyptus, a global tree for energy, paper and wood, 259–298 Springer New York. 2

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 7, 621–628 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 3, 14, 20, 22, 31, 34, 58, 73, 82, 83, 89, 141

Mungall, C. J., Emmert, D. B. and FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information *Bioinformatics* **23**, 13, i337–46. 38

Mungall, C. J., Misra, S., Berman, B. P., Carlson, J., Frise, E., Harris, N., Marshall, B., Shu, S., Kaminker, J. S., Prochnik, S. E., Smith, C. D., Smith, E., Tupy, J. L., Wiel, C., Rubin, G. M. and Lewis, S. E. (2002) An integrated computational pipeline and database to support whole-genome sequence annotation *Genome Biol* **3**, 12, RESEARCH0081. 42

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R., Usadel, B., Nikoloski, Z. and Persson, S. (2011) PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species *Plant Cell* **23**, 3, 895–910. 139

Myburg, A. A., Potts, B., Marques, C., Kirst, M., Gion, J., Grattapaglia, D. and Grima-Pettenati, J. (2005) *The genomes: a series on genome mapping, molecular breeding and genomics of economic species.* chapter Genome mapping and molecular breeding in Eucalyptus: molecular domestication of a major fiber crop. Enfield, NH, USA; Plymouth, UK: Science Publishers Inc. 2

Myers, E. W. (2005) The fragment assembly string graph *Bioinformatics* **21 Suppl 2**, ii79–85. 28

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 5881, 1344–1349 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 3, 14, 20

Ng, P., Tan, J. J. S., Ooi, H. S., Lee, Y. L., Chiu, K. P., Fullwood, M. J., Srinivasan, K. G., Perbost, C., Du, L., Sung, W.-K., Wei, C.-L. and Ruan, Y. (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes *Nucleic Acids Res* **34**, 12, e84. 19

Ng, P., Wei, C.-L., Sung, W.-K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., Liu, E. T. and Ruan, Y. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* **2**, 2, 105–111 ISSN 1548-7091 (Print); 1548-7091 (Linking). 5

Nielsen, K. L., Høgh, A. L. and Emmersen, J. (2006) DeepSAGE–digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples *Nucleic Acids Res* **34**, 19, e133. 21

Nishizawa, A., Yabuta, Y., Yoshida, E., Maruta, T., Yoshimura, K. and Shigeoka, S. (2006) Arabidopsis heat shock transcription factor A2 as a key regulator in response to several types of environmental stress *Plant J* **48**, 4, 535–47. 108

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Jr, Grattapaglia, D., Sederoff, R. R. and Kirst, M. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome *BMC Genomics* **9**, 312. 3, 19, 20, 82, 88, 118, 144

Nowrousian, M., Stajich, J. E., Chu, M., Engh, I., Espagne, E., Halliday, K., Kamerewerd, J., Kempken, F., Knab, B., Kuo, H.-C., Osiewacz, H. D., Poggeler, S., Read, N. D., Seiler, S., Smith, K. M., Zickler, D., Kuck, U. and Freitag, M. (2010) De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: Sordaria macrospora, a model organism for fungal morphogenesis. *PLoS Genet* **6**, 4, e1000891 ISSN 1553-7404 (Electronic); 1553-7390 (Linking). 15, 141

Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K. and Ohta, H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis *Nucleic Acids Res* **35**, Database issue, D863–9. 139

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes *Nucleic Acids Res* **27**, 1, 29–34. 48

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. and Li, P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 17, 3045–3054 ISSN 1367-4803 (Print); 1367-4803 (Linking). 35, 36, 44

Okoniewski, M. J. and Miller, C. J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* **7**, 276 ISSN 1471-2105 (Electronic);

1471-2105 (Linking). 20

Oracle (2009) A Comparison of Oracle Berkeley DB and Relational Database Management Systems White paper Oracle 500 Oracle Parkway, Redwood Shores, CA 94065, U.S.A. 85

Orvis, J., Crabtree, J., Galens, K., Gussman, A., Inman, J., Lee, E., Nampally, S., Riley, D., Sundaram, J., Felix, V., Whitty, B., Mahurkar, A., Wortman, J., White, O. and Angiuoli, S. (2010) Ergatis: A web interface and scalable software system for bioinformatics workflows. *Bioinformatics* ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 35, 38, 44

Oshlack, A. and Wakefield, M. J. (2009) Transcript length bias in RNA-seq data confounds systems biology *Biol Direct* **4**, 14. 80

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N. and Weigel, D. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads *Genome Res* **18**, 12, 2024–33. 19

Ozsolak, F., Platt, A., Jones, D., Reifenberger, J., Sass, L., McInerney, P., Thompson, J., Bowers, J., Jarosz, M. and Milos, P. (2009) Direct RNA sequencing *Nature*. 13

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 12, 1413–1415 ISSN 1546-1718 (Electronic); 1061-4036 (Linking). 14, 23, 34, 83

Pang, A., Macdonald, J., Pinto, D., Wei, J., Rafiq, M., Conrad, D., Park, H., Hurles, M., Lee, C., Venter, J., Kirkness, E., Levy, S., Feuk, L. and Scherer, S. (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**, 5, R52 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 14, 19

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H. and Soldatov, A. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA *Nucleic Acids Res.* 23, 24

Pavy, N., Pelgas, B., Beauseigle, S., Blais, S., Gagnon, F., Gosselin, I., Lamothe, M., Isabel, N. and Bousquet, J. (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce *BMC Genomics* **9**, 21. 18

Peleg, S., Sananbenesi, F., Zovoilis, A., Burkhardt, S., Bahari-Javan, S., Agis-Balboa, R. C., Cota, P., Wittnam, J. L., Gogol-Doering, A., Opitz, L., Salinas-Riester, G., Dettenhofer, M., Kang, H., Farinelli, L., Chen, W. and Fischer, A. (2010) Altered histone acetylation is associated with age-dependent memory impairment in mice *Science* **328**, 5979, 753–6. 38

Perkins, T. T., Kingsley, R. A., Fookes, M. C., Gardner, P. P., James, K. D., Yu, L., Assefa, S. A., He, M., Croucher, N. J., Pickard, D. J., Maskell, D. J., Parkhill, J., Choudhary, J., Thomson, N. R. and Dougan, G. (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS Genet* **5**, 7, e1000569 ISSN 1553-7404 (Electronic). 23, 24

Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly *Proc Natl Acad Sci U S A* **98**, 17, 9748–53. 28

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 7289, 768–772 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 24

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordonez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A. and Stratton, M. R. (2010*a*) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 7278, 191–196 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17, 141

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordonez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A. and Campbell, P. J. (2010*b*) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 7278, 184–190 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17, 141

Pushkarev, D., Neff, N. F. and Quake, S. R. (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**, 9, 847–852 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 13, 18

Ranik, M. and Myburg, A. A. (2006) Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis *Tree Physiol* **26**, 5, 545–56. 65

Rasmussen-Poblete, S., Valdes, J., Gamboa, M. C., Valenzuela, P. D. and Krauskopf, E. (2008) Generation and analysis of an Eucalyptus globulus cDNA library constructed from seedlings subjected to low temperature conditions *Electronic Journal of Biotechnology* **11**, 2 ISSN 0717-3458. 82

Reinhardt, J., Baltrus, D., Nishimura, M., Jeck, W., Jones, C. and Dangl, J. (2009) De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae *Genome research* **19**, 2, 294. 15, 141

Rengel, D., San Clemente, H., Servant, F., Ladouce, N., Paux, E., Wincker, P., Couloux, A., Sivadon, P. and Grima-Pettenati, J. (2009) A new genomic resource dedicated to wood formation in Eucalyptus *BMC Plant Biol* **9**, 36. 1, 82, 88

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 6, 276–277 ISSN 0168-9525 (Print); 0168-9525 (Linking). 87

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. and Pachter, L. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias *Genome Biol* **12**, 3, R22. 80

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A. and Birol, I. (2010) De novo assembly and analysis of RNA-seq data *Nat Methods.* 79

Rothberg, J. M. and Leamon, J. H. (2008) The development and impact of 454 sequencing. *Nature Biotechnology* **26**, 10, 1117–1124. 6

Royce, T. E., Rozowsky, J. S. and Gerstein, M. B. (2007) Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res* **35**, 15, e99 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 20

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallbook, F., Besnier, F., Carlborg, O., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K. and Andersson, L. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 7288, 587–591 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* **5**, 5, e1000386 ISSN 1553-7358 (Electronic). 30

Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W. and Velculescu, V. E. (2002) Using the transcriptome to annotate the genome *Nat Biotechnol* **20**, 5, 508–12. 21

Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. and Tettelin, H. (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 1, 24–31 ISSN 0888-7543 (Print); 0888-7543 (Linking). 87

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors *Proc Natl Acad Sci U S A* **74**, 12, 5463–7. 20

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C. and Jackson, S. A. (2010) Genome sequence of the palaeopolyploid soybean *Nature* **463**, 7278, 178–83. 16

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A. and Wilson, R. K. (2009) The B73 maize genome: complexity, diversity, and dynamics *Science* **326**, 5956, 1112–5. 16

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. and Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes *Genome Biol* **10**, 9, R98.

31

Schulze, U., Hepp, B., Ong, C. S. and Ratsch, G. (2007) PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics* **23**, 15, 1892–1900 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 34

Schuster, S. C. (2008) Next-generation sequencing transforms today's biology *Nat Methods* **5**, 1, 16–8. 4

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q., Burhans, R., Riemer, C., Wittekindt, N. E., Moorjani, P., Tindall, E. A., Danko, C. G., Teo, W. S., Buboltz, A. M., Zhang, Z., Ma, Q., Oosthuysen, A., Steenkamp, A. W., Oostuisen, H., Venter, P., Gajewski, J., Zhang, Y., Pugh, B. F., Makova, K. D., Nekrutenko, A., Mardis, E. R., Patterson, N., Pringle, T. H., Chiaromonte, F., Mullikin, J. C., Eichler, E. E., Hardison, R. C., Gibbs, R. A., Harkins, T. T. and Hayes, V. M. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 7283, 943–947 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A. and Shinozaki, K. (2002) Functional annotation of a full-length Arabidopsis cDNA collection *Science* **296**, 5565, 141–5. 82

Senger, M., Rice, P. and Oinn, T. (2003) SOAPlab - a unified Sesame door to analysis tools in *UK-eScience, All hands meeting* 509–513. 36

Shah, M. K., Lee, H., Rogers, S. A. and Touchman, J. W. (2004) An Exhaustive Genome Assembly Algorithm Using K-Mers to Indirectly Perform N-Squared Comparisons in O(N) in *CSB* 740–741. 28

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**, 10, 1135–1145 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 8, 13

Shendure, J., Mitra, R. D., Varma, C. and Church, G. M. (2004) Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**, 5, 335–344 ISSN 1471-0056 (Print); 1471-0056 (Linking). 4

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 5741, 1728–1732 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 4, 5, 8, 15, 141

Sibout, R., Eudes, A., Mouille, G., Pollet, B., Lapierre, C., Jouanin, L. and Séguin, A. (2005) CINNAMYL ALCOHOL DEHYDROGENASE-C and -D are the primary genes involved in lignin biosynthesis in the floral stem of Arabidopsis *Plant Cell* **17**, 7, 2059–76. 116

Simková, H., Safár, J., Suchánková, P., Kovárová, P., Bartos, J., Kubaláková, M., Janda, J., Cíhalíková, J., Mago, R., Lelley, T. and Dolezel, J. (2008*a*) A novel resource for genomics of Triticeae: BAC library specific for the short arm of rye (Secale cereale L.) chromosome 1R (1RS) *BMC Genomics* **9**, 237. 16

Simková, H., Svensson, J. T., Condamine, P., Hribová, E., Suchánková, P., Bhat, P. R., Bartos, J., Safár, J., Close, T. J. and Dolezel, J. (2008*b*) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley *BMC Genomics* **9**, 294. 16

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data *Genome Res* **19**, 6, 1117–23. 29

Singer, T., Fan, Y., Chang, H.-S., Zhu, T., Hazen, S. P. and Briggs, S. P. (2006) A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet* **2**, 9, e144 ISSN 1553-7404 (Electronic). 20

Sjödin, A., Street, N. R., Sandberg, G., Gustafsson, P. and Jansson, S. (2009) The Populus Genome Integrative Explorer (PopGenIE): a new resource for exploring the Populus genome *New Phytol* **182**, 4, 1013–25. 123, 144

Slater, G. S. C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison *BMC Bioinformatics* **6**, 31. 46, 49, 74

Smith, A. D., Xuan, Z. and Zhang, M. Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping *BMC Bioinformatics* **9**, 128. 31

Srivastava, S. and Chen, L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data *Nucleic Acids Res* **38**, 17, e170. 80

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. and Birney, E. (2002) The Bioperl toolkit: Perl modules for the life sciences *Genome Res* **12**, 10, 1611–8. 43

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215–25 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 87

Stein, L. (2002) Creating a bioinformatics nation *Nature* **417**, 119–120. 36

Stein, L. D. (2010) The case for cloud computing in genome informatics *Genome Biol* **11**, 5, 207. 34, 36

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. and Lewis, S. (2002) The generic genome browser: a building block for a model organism system database *Genome Res* **12**, 10, 1599–610. 38

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., Van Montagu, M., Sandberg, G., Olsson, O., Teeri, T. T., Boerjan, W., Gustafsson, P., Uhlén, M., Sundberg, B. and Lundeberg, J. (1998) Gene discovery in the wood-forming tissues of poplar: analysis of 5, 692 expressed sequence tags *Proc Natl Acad Sci U S A* **95**, 22, 13330–5. 82

Steurnagel, B., Taudien, S., Gundlach, H., Seidel, M., Ariyadasa, R., Schulte, D., Petzold, A., Felder, M., Graner, A., Scholz, U., Mayer, K. F. X., Platzer, M. and Stein, N. (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley *BMC Genomics* **10**, 547. 15, 16

Stromberg, M. and Marth, G. (2008) MOSAIK: A reference-guided assembler for next-generation sequence data *Manuscript in preparation.* 30, 31, 85, 86

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. and Yaspo, M.-L. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 5891, 956–960 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 14, 23, 83

Sun, Z., Asmann, Y. W., Kalari, K. R., Bot, B., Eckel-Passow, J. E., Baker, T. R., Carr, J. M., Khrebtukova, I., Luo, S., Zhang, L., Schroth, G. P., Perez, E. A. and Thompson, E. A. (2011) Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing *PLoS One* **6**, 2, e17490. 141

Swaminathan, K., Alabady, M. S., Varala, K., De Paoli, E., Ho, I., Rokhsar, D. S., Arumuganathan, A. K., Ming, R., Green, P. J., Meyers, B. C., Moose, S. P. and Hudson, M. E. (2010) Genomic and small RNA sequencing of Miscanthus x giganteus shows the utility of sorghum as a reference genome sequence for Andropogoneae grasses *Genome Biol* **11**, 2, R12. 16

Tan, W., Foster, I. and Madduri, R. (2008) Combining the Power of Taverna and caGrid: Scientific Workflows that Enable Web-Scale Collaboration *IEEE Internet Computing* **12**, 61–68. 37

Tan, W., Missier, P., Foster, I., Madduri, R., De Roure, D. and Goble, C. (2009) A comparison of using Taverna and BPEL in building scientific workflows: the case of caGrid *Concurrency and Computation: Practice and Experience*. 37

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. and Surani, M. A. (2009) mRNA-Seq whole-transcriptome analysis of a single cell *Nat Methods* **6**, 5, 377–82. 80

Tauch, A., Trost, E., Tilker, A., Ludewig, U., Schneiker, S., Goesmann, A., Arnold, W., Bekel, T., Brinkrolf, K., Brune, I., Götker, S., Kalinowski, J., Kamp, P.-B., Lobo, F. P., Viehoever, P., Weisshaar, B., Soriano, F., Dröge, M. and Pühler, A. (2008) The lifestyle of Corynebacterium urealyticum derived from its complete genome sequence established by pyrosequencing *J Biotechnol* **136**, 1-2, 11–21. 15, 141

Tawfik, D. S. and Griffiths, A. D. (1998) Man-made cell-like compartments for molecular evolution. *Nat*

*Biotechnol* **16**, 7, 652–656 ISSN 1087-0156 (Print); 1087-0156 (Linking). 5

Taylor, J., Schenck, I., Blankenberg, D. and Nekrutenko, A. (2007) Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics* **Chapter 10**, Unit 10.5 ISSN 1934-340X (Electronic); 1934-3396 (Linking). 43

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 9, 1105–1111 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 31, 34, 48, 74, 80

Trapnell, C. and Salzberg, S. L. (2009) How to map billions of short reads onto genomes. *Nat Biotechnol* **27**, 5, 455–457 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 29, 104

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 5, 511–515 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 23, 49, 73, 80, 83, 89, 104, 130, 143

Travers, K., Chin, C., Rank, D., Eid, J. and Turner, S. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 11, 12

Turcatti, G., Romieu, A., Fedurco, M. and Tairi, A.-P. (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res* **36**, 4, e25 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 4, 6, 7

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y. and Rokhsar, D. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 5793, 1596–1604 ISSN 1095-9203 (Electronic); 1095-9203 (Linking). 2, 16, 82, 88

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. and Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data *Nat Methods* **5**, 9, 829–34. 141

van Bakel, H., Nislow, C., Blencowe, B. J. and Hughes, T. R. (2010) Most "dark matter" transcripts are associated with known genes *PLoS Biol* **8**, 5, e1000371. 22

van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Verstegen, H. and van Eijk, M. J. T. (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes *PLoS One* **2**, 11, e1172. 19

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C. and Sonstegard, T. S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5**, 3, 247–252 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 16

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L. M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J. T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Demattè, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J. A., Sterck, L., Vandepoele, K., Grando, S. M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S. K., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F. and Viola, R. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety *PLoS One* **2**, 12, e1326. 16

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 5235, 484–487 ISSN 0036-8075 (Print); 0036-8075 (Linking). 20

Velicer, G. J., Raddatz, G., Keller, H., Deiss, S., Lanz, C., Dinkelacker, I. and Schuster, S. C. (2006) Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proc Natl Acad Sci U S A* **103**, 21, 8107–8112 ISSN 0027-8424 (Print); 0027-8424 (Linking). 15, 16

Venter, J. and Smith, H. (1996) A new strategy for genome sequencing. *Nature*. 15

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001) The sequence of the human genome *Science* **291**, 5507, 1304–51. 3, 15, 18

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I. and Marden, J. H. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing *Mol Ecol* **17**, 7, 1636–47. 83

Vizoso, P., Meisel, L. A., Tittarelli, A., Latorre, M., Saba, J., Caroca, R., Maldonado, J., Cambiazo, V., Campos-Vargas, R., Gonzalez, M., Orellana, A. and Silva, H. (2009) Comparative EST transcript profiling of peach fruits under different post-harvest conditions reveals candidate genes associated with peach fruit quality *BMC Genomics* **10**, 423. 79

Wang, B.-B. and Brendel, V. (2006) Genomewide comparative analysis of alternative splicing in plants *Proc Natl Acad Sci U S A* **103**, 18, 7175–80. 23

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K.-S., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H. and Wang, J. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**, 7218, 60–65 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18

Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010*a*) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 1, 136–138 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 46, 49, 76, 80, 143

Wang, L., Li, P. and Brutnell, T. P. (2010*b*) Exploring plant transcriptomes using ultra high-throughput sequencing *Briefings in Functional Genomics* **9**, 2, 1–11. 141

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 1, 57–63 ISSN 1471-0064 (Electronic); 1471-0056 (Linking). 20

Wanner, L. A., Li, G., Ware, D., Somssich, I. E. and Davis, K. R. (1995) The phenylalanine ammonia-lyase gene family in Arabidopsis thaliana *Plant Mol Biol* **27**, 2, 327–38. 108

Wehmeyer, N. and Vierling, E. (2000) The expression of small heat shock proteins in seeds responds to discrete developmental signals and suggests a general protective role in desiccation tolerance *Plant Physiol* **122**, 4, 1099–108. 108

Werner, T. (2010) Next generation sequencing in functional genomics. *Brief Bioinform* ISSN 1477-4054 (Electronic); 1467-5463 (Linking). 3

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A. and Rothberg, J. M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 7189, 872–876 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18

Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. and Stein, N. (2006) 454 sequencing put to the test using the complex genome of barley *BMC Genomics* **7**, 275. 15

Wilhelm, B. T. and Landry, J.-R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 3, 249–257 ISSN 1095-9130 (Electronic); 1046-2023 (Linking). 21, 141

Wilkinson, M. D. and Links, M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform* **3**, 4, 331–341 ISSN 1467-5463 (Print); 1467-5463 (Linking). 36

Wittkopp, P. J., Haerum, B. K. and Clark, A. G. (2008) Independent effects of cis- and trans-regulatory variation on gene expression in Drosophila melanogaster *Genetics* **178**, 3, 1831–5. 145

Wu, T., Qin, Z., Zhou, X., Feng, Z. and Du, Y. (2010) Transcriptome profile analysis of floral sex determination in cucumber *J Plant Physiol* **167**, 11, 905–13. 83

Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., Dai, F., Li, Y., Cheng, D., Li, R., Cheng, T., Jiang, T., Becquet, C., Xu, X., Liu, C., Zha, X., Fan, W., Lin, Y., Shen, Y., Jiang, L., Jensen, J., Hellmann, I., Tang, S., Zhao, P., Xu, H., Yu, C., Zhang, G., Li, J., Cao, J., Liu, S., He, N., Zhou, Y., Liu, H., Zhao, J., Ye, C., Du, Z., Pan, G., Zhao, A., Shao, H., Zeng, W., Wu, P., Li, C., Pan, M., Li, J., Yin, X., Li, D., Wang, J., Zheng, H., Wang, W., Zhang, X., Li, S., Yang, H., Lu, C., Nielsen, R., Zhou, Z., Wang, J., Xiang, Z. and Wang, J. (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* **326**, 5951, 433–436 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 17

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L. and Yang, H. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica) *Science* **296**, 5565, 79–92. 16

Yuen, C. Y. L., Pearlman, R. S., Silo-Suh, L., Hilson, P., Carroll, K. L. and Masson, P. H. (2003) WVD2 and WDL1 modulate helical organ growth and anisotropic cell expansion in Arabidopsis *Plant Physiol* **131**, 2, 493–506. 108

Zdobnov, E. M. and Apweiler, R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro *Bioinformatics* **17**, 9, 847–8. 46, 48, 70, 143

Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 5, 821–829 ISSN 1088-9051 (Print). 28, 29, 46, 56, 64, 84, 142

Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., Wei, S., Fu, J., Chen, Y., Ren, X., Zhang, Y., Ni, P., Zhang, J., Li, S., Wang, J., Wong, G. K.-S., Zhao, H., Yu, J., Yang, H. and Wang, J. (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics *Nucleic Acids Res* **32**, Database issue, D377–82. 123

Zheng, M. S., Takahashi, H., Miyazaki, A., Hamamoto, H., Shah, J., Yamaguchi, I. and Kusano, T. (2004) Up-regulation of Arabidopsis thaliana NHL10 in the hypersensitive response to Cucumber mosaic virus infection and in senescing leaves is controlled by signalling pathways that differ in salicylate involvement *Planta* **218**, 5, 740–50. 108