UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA

Vid JELEN

PRIMERJAVA MITOHONDRIJSKIH GENOMOV IN ANALIZA MUTACIJ JEDRNIH GENOV ŠESTIH SEVOV FITOPATOGENE GLIVE *Verticillium albo-atrum*

DOKTORSKA DISERTACIJA

Ljubljana, 2017

UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA

Vid JELEN

PRIMERJAVA MITOHONDRIJSKIH GENOMOV IN ANALIZA MUTACIJ JEDRNIH GENOV ŠESTIH SEVOV FITOPATOGENE GLIVE *Verticillium albo-atrum*

DOKTORSKA DISERTACIJA

MITOCHONDRIAL AND NUCLEAR GENOME COMPARISON OF SIX *Verticillium albo-atrum* PHYTOPATHOGENIC FUNGI STRAINS

DOCTORAL DISSERTATION

Ljubljana, 2017

Na podlagi Statuta Univerze v Ljubljani ter po sklepu Senata Biotehniške fakultete in sklepa Komisije za doktorski študij Univerze v Ljubljani z dne 13.11.2013 (po pooblastilu Senata Univerze z dne 20.1.2009) je bilo potrjeno, da kandidat izpolnjuje pogoje za opravljanje doktorata znanosti na Interdisciplinarnem doktorskem študijskem programu Bioznanosti, znanstveno področje: bioinformatika. Za mentorja je bil imenovan izr. prof. dr. Jernej Jakše.

Doktorska disertacija je zaključek Interdisciplinarnega doktorskega študijskega programa bioznanosti, področje bioinformatika. Delo je bilo opravljeno v laboratorijih Oddelka za agronomijo, Biotehniške Fakultete, Univerze v Ljubljani ter v prostorih skupine Bioinformatics and Evolutionary Genomics group z oddelka Plant Systems Biology na Flamskem inštitutu za biotehnologijo, Gent, Belgija.

Komisija za oceno in zagovor:

Predsednik: prof. dr. Peter DOVČ
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za zootehniko

Član: prof. dr. Gregor ANDERLUH
Kemijski inštitut, Laboratorij za molekularno biologijo in nanobiotehnologijo

Član: doc. dr. Tomaž CURK
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Datum zagovora:

Podpisani izjavljam, da je disertacija rezultat lastnega raziskovalnega dela. Izjavljam, da je elektronski izvod identičen tiskanemu. Na Univerzo neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravici shranitve avtorskega dela v elektronski obliki in reproduciranja ter pravico omogočanja javnega dostopa do avtorskega dela na svetovnem spletu preko Digitalne knjižnice Biotehniške fakultete.

Doktorand:

Vid JELEN

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

ŠD Dd
DK UDK 577.2: 575.112(043)=163.6
KG bioinformatika/primerjalna genomika/Verticillium/mitohondriji/evolucija/filogenetska analiza/mutacijska analiza/mitohondrijski genom/sekvenciranje naslednje generacije
AV JELEN, Vid, univ. dipl. bioteh.
SA JAKŠE, Jernej (mentor)
KZ SI-1000 Ljubljana, Jamnikarjeva 101
ZA Univerza v Ljubljani, Biotehniška fakulteta, Interdisciplinarni doktorski študij Bioznanosti, znanstveno področje bioinformatika
LI 2017
IN PRIMERJAVA MITOHONDRIJSKIH GENOMOV IN ANALIZA MUTACIJ JEDRNIH GENOV ŠESTIH SEVOV FITOPATOGENE GLIVE *Verticillium albo-atrum*
TD Doktorska disertacija
OP XIII, 105 str. 17 pregl., 41 sl., 9 pril., 103 vir.
IJ sl
JI sl/en
AI Z uporabo trenutno aktualnih bioinformatičnih metod in pristopov znanstvenega programiranja smo analizirali mitohondrijski genom ter jedrne gene šestih sevov fitopatogene glive *V. albo-atrum*, ki primarno okužujejo hmelj. Uspeli smo sestaviti in pripisati genomske značilnosti 26.139 bp velikemu referenčnemu mitohondrijskemu genomu, ki vsebuje 14 protein-kodirajočih genov, rRNA podenote, ribosomalni protein S3 in 26 tRNA-kodirajočih genov. Poleg teh značilnosti, ki so pogosta lastnost mitohondrijskih genomov gliv iz taksonomskega razreda *Sordariomycetes,* smo odkrili tudi potencialno dolgo ne-kodirajočo RNA (*orf414*). Filogenetska analiza našega mitohondrijskega genoma je pokazala, da ta spada v skupino *Glomerellales* v razredu *Sordariomycetes* in da je najbližji sorodnik *V. dahliae.* Ob vpogledu v jedrni genom glive smo preučili gostoto porazdelitev variant, eksonskih ter ponovitvenih regij in na podlagi variant s pripisanimi značilnostmi pripravili filogenetsko drevo odnosov med sevi. To drevo je prikazalo jasno ločitev letalnega in blagega patotipa *V. albo-atrum,* vendar ni omogočilo tudi določiti izvornega seva oz. korenine drevesa. Pri analizi mutacij jedrnih genov smo se poslužili metode določitve evolucijskega pritiska s Ka/Ks koeficienti. Ob tem smo uporabili 2 metodi izračuna koeficienta - Nei-Gojobori ter Yang-Nielsen. Z njima smo določili skupine genov, ki so pod evolucijskim pritiskom glede na primerjave z drugimi vrstami rodu *Verticillium* in jim na koncu določili še obogatenost GO pojmov. Poseben poudarek pri zasnovi analiz je bila tudi njihova ponovljivost in preprostost izvajanja, za kar smo poskrbeli z razvojem lastnega programskega ogrodja za izvajanje analiz.

KEY WORDS DOCUMENTATION

DN Dd
DC UDC 577.2: 575.122(043)=163.6
CX bioinformatics/comparative genomics/Verticillium/mitochondrion/evolution/ phylogenetic analysis/mutational analysis/next generation sequencing
AU JELEN, Vid
AA JAKŠE, Jernej (supervisor)
PP SI-1000 Ljubljana, Jamnikarjeva 101
PB University of Ljubljana, Biotechnical Faculty, Interdisciplinary Doctoral Programme in Biosciences, Scientific Field Bioinformatics
PY 2017
TI MITOCHONDRIAL AND NUCLEAR GENOME COMPARISON OF SIX *Verticillium albo-atrum* PHYTOPATHOGENIC FUNGI STRAINS
DT Doctoral Dissertation
NO XIII, 105 p., 17 tab., 41 fig., 9 ann., 103 ref.
LA sl
AL sl/en
AB By applying currently up-to-date bioinformatics methods and scientific programming approaches, we have analyzed the mitochondrial genome and nuclear genes of six *V. albo-atrum* phytopathogenic fungi strains, which primarily infect hop plants. The former analysis resulted in a 26.139 bp reference mitochondrial genome, encompassing 14 protein-coding genes, rRNA sub-units, a ribosomal protein S3 and 26 tRNA-coding genes. Besides those features we also uncovered a potential long non-coding RNA (*orf414*). A phylogenetic analysis of this mitochondrial genome has shown, that it clusters along with other members of the *Glomerellales* group of *Sordariomycetes* and that it is most closely related to *V. dahliae.* During a general overview of the genome we studied the density distributions of its variants, exonic and repetitive regions and we prepared a phylogenetic tree of strain relations from the annotated variants coming from their genomes. This tree has shown a clear distinction between the mild and lethal *V. albo-atrum* pathotype, but did not also result in a source strain or the tree root. By analyzing mutations in the nuclear genes, we applied methods of determining the evolutionary pressure with Ka/Ks coefficients. Utilizing this approach we used 2 methods of coefficients calculation - NG and YN. By adopting these methods, we determined the groups of genes under evolutionary pressure with comparisons to other species. A GO enrichment analysis followed after the groups were established. A special emphasis at analysis creation was put on their reproducibility and simplicity of execution, which we addressed by developing a custom framework for executing analyses.

KAZALO VSEBINE

KAZALO PREGLEDNIC

KAZALO SLIK

KAZALO PRILOG

OKRAJŠAVE IN SIMBOLI

POJEM	*RAZLAGA*	*ANGLEŠKI IZRAZ*
ATP	adenozin tri fosfat	Adenosine tri-phosphate
BAM	binarni zapis SAM formata	Binary SAM format
cpDNA	kloroplastna DNA	Chloroplast DNA
COX	ciklooksigenaza	Cyclooxygenase
DEL	izbris	Deletion
ENA	evropski nukleotidni arhiv	European nucleotide archive
EST	oznaka izraženega zaporedja	Expressed sequence tag
FL-cDNA	cDNA celotne dolžine	Full-length cDNA
HMM	skriti markovski model	Hidden markov model
INS	vključitev	Insertion
MNP	polimorfizem več nukleotidov	Multiple nucleotide polymorphism
mtDNA	mithondrijska DNA	Mitochondrial DNA
NADH	nikotinamid dinukleotid hidrogenaza	Nicotinamide dinucleotide hydrogenase
NGS	sekvenciranje naslednje generacije	Next generation sequencing
ORF	odprti bralni okvir	Open reading frame
RNL	velika podenota rRNA	Large rRNA subunit
RNS	majhna podenota rRNA	Small rRNA subunit
SAM	tekstovni format zapisa poravnave/kartiranja zaporedij	Sequence alignment/map format
SNP	polimorfizem posameznega nukleotida	Single nucleotide polymorphism
SVM	metoda podpornih vektorjev	Support vector machine
UTR	neprevedena regija	Untranslated region
WGS	hitri postopek sekvenciranja	Whole-genome shotgun

SLOVARČEK

Kontig - Neprekinjen linearen niz DNA ali RNA zaporedja. Lahko je sestavljen tudi iz več manjših, delno prekrivajočih fragmentov - odčitkov.

K-mera - Kratek, unikaten element DNA dolžine "*k*", ki se uporablja pri mnogih algoritmih za sestavljanje zaporedij.

Odčitek - Kratko DNA ali RNA zaporedje, pridobljeno iz matičnega zaporedja z uporabo visoko-zmogljivega sekvenciranja.

Ogrodje - Združek odčitkov na podlagi podobnosti v večje sestavke.

Primerjalna genomika – Področje, ki se osredotoča na primerjavo genomskih lastnosti med različnimi organizmi za preučevanje podobnosti in razlik v strukturah in evoluciji.

Programski cevovod (pipeline) - Zaporedje programskih procesov, ki so nanizani tako, da rezultat nekega procesa predstavlja vhodni podatek naslednjemu procesu.

Pokritost zaporedja - Povprečno število odčitkov na posameznem lokusu.

Poravnava zaporedij - Poravnava bioloških zaporedij, osnovana na njihovi medsebojni podobnosti.

1 UVOD

Glivni fitopatogeni so povzročitelji številnih rastlinskih bolezni. Okužbe z njimi lahko izrazito vplivajo na količino in kakovost pridelka ter posledično povzročijo tudi obsežnejše proizvodne in finančne posledice v komercialni proizvodnji rastlin. Za učinkovite implementacije strategij njihovega zatiranja je potrebno stalno izboljševati in razvijati nove načine obvladovanja virulentnosti teh patogenov, kar pa je mogoče le z dobrim poznavanjem njihovih mehanizmov patogenosti in virulentnih dejavnikov. Predvsem pomembno področje pri tem je preučevanje molekularnih mehanizmov, ki so specifični za posamezni fitopatogeni organizem. Zaradi velike količine možnih podatkov o molekularni sestavi patogena in njegovega genetskega ustroja to področje predstavlja izdaten izziv. Za analizo te velike količine raznolikih podatkov pa je najprimernejša aplikacija pristopov s področja bioinformatike, ki združuje računalniške vede s statističnimi analizami za obdelavo bioloških podatkov.

Možnost določanja nukleotidnega zaporedja celotnih genomov s pomočjo tehnologij sekvenciranja nove generacije (NGS) je sprožila znanstveno revolucijo v bioinformatiki, pri čemer je prišlo do poplave ogromnih količin podatkov in potrebe po novih pristopih za njihovo zajemanje in obdelavo (Schadt in sod., 2010). Tehnološki napredek je privedel do izjemnega števila objav s tega področja, hkrati pa je omogočil nove pristope s celostnim preučevanjem bioloških informacij, ki jih nukleotidna zaporedja vsebujejo. Vzporedno z razvojem NGS tehnologij je prišlo tudi do razvoja na področju računalništva; povečale so se tako sposobnosti strojne kot tudi programske opreme za shranjevanje in procesiranje bioloških podatkov. Za analize te množice sekvenčnih podatkov je na voljo več različnih pristopov: npr. uporaba lokalne računalniške opreme, uporaba strežnika znotraj omrežja, uporaba storitev računalništva v oblaku. Programska oprema, ki jo uporabimo za analizo bioloških podatkov, je lahko plod lastnega programiranja ali pa rezultat drugega raziskovalca oz. skupine, ki jo ponudi kot odprto kodno različico, prosto dostopno na spletu ali pa tudi kot plačljivo programsko opremo.

Splošna dostopnost bioloških zaporedij preko javno dostopnih podatkovnih baz in uporaba zmogljive računalniške opreme ter novih algoritmov obdelave sta omogočila preučevanje organizmov na različnih nivojih, kot npr.: genoma (celotnega nabora DNA molekul znotraj posamezne celice organizma), transkriptoma (celotnega nabora izraženih genov) ali točno določenih tarčnih zaporedij (Chip-seq, Cot-filtracija, metilacijska filtracija itd.). S sekvenciranjem, sestavljanjem, analizo struktur in primerjavo genomov lahko na podlagi podobnosti in razlik funkcionalnih elementov (npr. genov, proteinov, regulatornih regij, …) sklepamo, kakšne so njihove funkcije in kako je selekcija delovala nanje.

Veliko število metod za napovedovanje funkcij genomskih elementov temelji na identifikaciji, karakterizaciji in kvantifikaciji podobnosti med preučevanim elementom in elementi, za katere so podatki o njihovih funkcijah na voljo. Ker je zaporedje glavni dejavnik funkcije, se na podlagi podobnosti zaporedij implicitno sklepa na podobnost funkcij teh zaporedij. En pristop k odkritju funkcij zaporedij, ki je pogost na področju strojnega učenja, je zbiranje učne množice poravnav zaporedij z znano funkcijo, kot npr. regulatorna zaporedja, in uporaba teh poravnav za razvoj statističnih modelov (kot npr. skritih markovskih modelov), ki ocenijo verjetnost, da je lahko dano poravnavo generiral model. Obenem lahko s primerjavo sestavljenih genomov odkrivamo zaporedja, ki so ohranjena med vrstami, kot tudi zaporedja, ki so značilna za posamezne organizme. DNA zaporedja, ki kodirajo proteine in RNA molekule s podobno funkcijo v različnih organizmih, bodo imela visoko stopnjo ohranjenosti, medtem ko bodo ostala zaporedja kazala znake razhajanja.

Bioinformatične metode, ki smo jih uporabili v tem doktorskem delu, so v veliki meri implementirane kot prosto dostopna programska oprema, ki se splošno uporablja v genomskih raziskavah. Kjer se je pokazala potreba za rešitev po meri, smo razvili tudi lastne programske skripte. Pri analizah smo poleg lastnih eksperimentalnih podatkov uporabili tudi podatke iz javno dostopnih podatkovnih baz. Ti so nam omogočili sintezo podatkov na več nivojih - genoma, transkriptoma in proteoma, s čimer smo poskusili uporabiti čimveč razpoložljivih informacij za naše raziskave. Predmet raziskav v tem doktorskem delu je bilo preučevanje genomskih lastnosti, ki so odraz evolucijskega razvoja fitopatogene glive *Verticillium albo-atrum* v jedrnem in mitohondrijskem genomu.

1.1 NAMEN RAZISKAVE

V raziskavi smo predpostavili, da se bo s pomočjo genomskih podatkov sevov glive *Verticillium albo-atrum* REC (Slovenija, patotip PG1), T2 (Slovenija, patotip PG2), 1953 (Anglija, patotip PG1), 1985 (Anglija, patotip PG2), P55 (Nemčija, patotip PG1) in P15 (Nemčija, patotip PG2), pridobljenih z NGS tehnologijami in bioinformacijskimi orodji:

- Za vseh šest sevov glive preverilo uspešnost *de-novo* sestavljanja mitohondrijskih zaporedij, opravil pripis genomskih značilnosti in primerjalna analiza sestavljenih mitohondrijskih genomov. **Hipoteza:** zaradi konzervativne narave mitohondrijskih genomov ne pričakujemo večjih razlik med mitohondrijskimi genomi šestih sevov ali patotipsko specifičnih razlik. Preučile se bodo tudi morebitne razlike z mitohondrijskimi genomi ostalih vrst za namen potencialnih biomarkerjev.
- Za vseh šest sevov določilo stopnje mutacij jedrnih genov v primerjavi z referenčnim genomom, ki bodo pomagale razumeti, kako se genomi razlikujejo glede na referenčni genom in predpostavilo selekcijske scenarije. **Hipoteza:** pričakujemo potrditev regij genoma z različnimi stopnjami mutacij in regije genoma, kjer bodo zaznani značilno različni selekcijski pritiski med patotipoma.
- Preverilo ali se dejavniki, ki vodijo do povečane virulentnosti sevov, spreminjajo pod vplivom pozitivne selekcije. **Hipoteza:** pozitivna selekcija je povzročila pospešen razvoj virulentnih dejavnikov

1.1.1 Sklop 1 – Priprava in filogenetska analiza mitohondrijskega genoma

Mitohondrijska DNA ni metilirana, vsebuje dobro ohranjene temeljne protein-kodirajoče gene in je v vsaki celici prisotna v več kopijah. Zaradi teh lastnosti se zaporedja mitohondrijskih genov pogosto uporabljajo kot markerji za populacijske in vrstne karakterizacije (Torriani in sod., 2014). V okviru naše raziskave smo se zato odločili pripraviti mitohondrijski genom *V. albo-atrum* in ga uporabiti v filogenetski analizi, da bi preverili razmerja med našimi preučevanimi sevi. Obenem pa smo sklenili rezultate te analize primerjati tudi z že opravljenimi filogenetskimi analizami na podlagi jedrnih genomov, s čimer smo želeli preveriti domnevo, da mitohondrijska DNA in jedrna DNA odsevata enako sliko filogenetskih odnosov, oziroma so razlike med njima minimalne.

1.1.2 Sklop 2 – Genomska filogenija in mutacijske analize

Pri pridelavi hmelja so se dosedaj pojavile letalne oblike *Verticillium albo-atrum* na treh različnih geografskih področjih: Anglija - 1933 (Keyworth, 1942), Slovenija - 1997 (Radišek in sod., 2003) in Nemčija - 2005 (Seefelder in sod., 2009). Naša hipoteza je, da so se letalne oblike *Verticillium albo-atrum* razvile iz prilagoditve blagih oblik na specifične gostitelje. Za preverjanje te hipoteze smo se odločili uporabiti 2 pristopa: analizo filogenetskih odnosov teh sevov glede na njihove genomske variante ter analize evolucijskega pritiska na kodirajoča zaporedja v njihovih genomih preko Ka/Ks koeficientov. Druga domneva, ki smo jo tudi želeli preveriti pa je, ali so se letalni sevi v vsaki državi razvili neodvisno drug od drugega, ali so to le potomci enega seva, ki je razvil letalni fenotip.

1.1.3 Genom *V. albo-atrum*

Referenčni genom s pripisanimi genomskimi značilnostmi *V. albo-atrum* T2 – letalni sev, je bil pripravljen na katedri za agronomijo že pred začetkom tega dela, kot tudi kartiranje odčitkov ostalih sevov na ta genom, zato je nadaljnje delo temeljilo na nekaterih vnaprej pripravljenih podatkih. Referenčni genom je velik 31,3 Mb z 9.269 genskimi modeli in vsebuje 550 kb DNA zaporedij, ki so specifične za letalni patotip in 90 genskih modelov, ki se nahajajo znotraj teh sekvenc. Regije, ki so patotipno specifične za letalni tip, so bile analizirane in funkcijsko preučene v drugem projektu in jih zato v tem doktorskem delu ne obravnavamo.

1.2 CILJI

Za pripravo genomskih virov glive *V. albo-atrum* se je izvedlo sekvenciranje celotnih genomov in RNA-Seq prej omenjenih sevov. Na podlagi teh podatkov smo v skupnem raziskali 3 letalne in 3 blage seve, pri čemer je bil po en par blag-letalni iz vsake prej omenjene geografske regije. Z uporabo sodobnih bioinformatičnih metod in pristopov pri analizi pridobljenih podatkov se je preučilo genomske značilnosti obravnavanih sevov fitopatogene glive *V. albo-atrum.* Namen teh raziskav je bil primerjati mitohondrijske genome med sevi ter preučiti stopnjo mutacij jedrnih genov referenčnega seva na podlagi evolucijskega pritiska in na podlagi teh ugotovitev pripraviti morebitne selekcijske scenarije, ki so privedli do nastanka sevov s povišano patogenostjo (″letalnega″ patotipa).

2 PREGLED OBJAV

Ob začetku raziskav tega doktorskega dela je bila naša preučevana fitopagena gliva klasificirana kot *Verticillium albo-atrum*. Po tej prvotni klasifikaciji je bila vrsta *V. albo-atrum* razdeljena na skupini Grp I in Grp II, pri čemer se je prva skupina nadalje delila na dve podskupini. Razlika med njima je bila v patogenosti do lucerne (*V. alfalfae in V. nonalfalfae*). Na podlagi novih dognanj (Inderbitzin in sod., 2011) sta se ti podskupini določili kot samostojni vrsti in tako se je uveljavila novejša klasifikacija našega preučevanega organizma kot *Verticillium nonalfalfae,* ki bo nadalje splošno uporabljena v besedilu (Inderbitzin in sod., 2011).

2.1 VRSTA *Verticillium nonalfalfae*

Verticillium nonalfalfae je glivni rastlinski patogen, ki se nahaja v zemlji in okužuje več kot 400 različnih rastlinskih vrst. Ekonomsko najpomembnejša gostitelja sta hmelj in bombaž. Gliva ob okužbi gostitelja povzroči rumenenje in venenje rastline, kar izdatno vpliva na stanje, donos ter posledično tudi ekonomsko vrednost pridelka (Radišek in sod., 2003).

Obstaja več različnih sevov tega patogena, med katerimi je prisotna velika variabilnost v virulenci in naboru gostiteljev. Te po ustaljeni klasifikaciji v grobem razdelimo na »blag« in »letalni« patotip (ali tudi »blage« in »letalne« seve) *Verticillium nonalfalfae*. Simptomi blage verticilijeve uvelosti precej nihajo v razsežnosti iz leta v leto v okuženem hmeljnem nasadu glede na temperaturo zemlje, razpoložljivost dušika in drugih okoljskih dejavnikov. Rastline, okužene z blago obliko te bolezni, se v večini primerov pozdravijo in v naslednjem letu ne kažejo več simptomov okužbe. Nasprotno od te oblike, letalna oblika verticilijeve uvelosti vodi v uvelost celotne rastline in povzroči odmrtje pri dovzetnih kultivarjih. Okoljske razmere in dejavniki v zemlji nimajo velikega vpliva na pojavnost te oblike bolezni (Gent in sod., 2012).

2.2 GENOMIKA RASTLINSKIH GLIVNIH PATOGENOV

Glive so povzročiteljice mnogih resnih rastlinskih bolezni in imajo med mikrobnimi patogeni to unikatno lastnost, da lahko prestopijo vrhnji zaščitni sloj rastline in s tem hitro povzročijo okužbo, kar pogosto privede do velikih posledic v komercialni produkciji rastlin. Njihova patogenost je kompleksen fenotip, ker predstavlja sposobnost razvoja infektivnih struktur, invazijo živih rastlinskih tkiv ter tvorbo struktur za

nadaljnje širjenje bolezni, poleg perturbacije gostiteljske celične signalizacije in metabolizma, kar je značilno za bakterijsko patogenezo (Soanes in sod., 2007).

Nekateri filamentozni glivni patogeni so se razvili s prehodom z ene gostiteljske rastline na drugo, pri čemer so se pogosto premaknili na rastlino, ki je v daljnem sorodstvu s prvotnim gostiteljem. Takšni dogodki so pogosto izjemno vplivali na evolucijo njihovih efektorjev - modulatorjev procesov v gostiteljski rastlini. V nekaterih primerih so se izvorni efektorji prilagodili na nove tarče preko nabiranja mutacij, ki so izboljšale ali razširile aktivnost njihovih efektorjev, v primerih velikih razlik med gostitelji pa je lahko podmnožica efektorjev postala celo odvečna, ker njihove tarče niso bile na voljo v novem gostitelju in so se tekom evolucije ti efektorji odstranili iz genoma. Primer takšne delecije efektorskih genov se je eksperimentalno dokazal v genomu *M. pennsylvanicum*, ki se je domnevno iz patogena trav razvil v patogena dvokaličnic (Dong in sod., 2015). Za pridobitev novih virulenčnih funkcij je lahko odgovoren tudi lateralen prenos genov, katerega je možno zaznati s primerjavo genomov patogenov (Soanes in sod., 2007).

V zadnjih 10 letih se je s pomočjo sekvenciranja in analize genomov filamentoznih glivnih patogenov predvidelo nekaj skupnih temeljnih lastnosti. Najizrazitejša izmed njih je ta, da vsebujejo velike nabore genov za virulenčne efektorje. Ti geni niso naključno razporejeni po genomu, ampak se največkrat združujejo v predele, bogate z zaporedji ponovitev in transpozabilnih elementov. Na podlagi teh ugotovitev se je predpostavil model genoma dveh hitrosti, kjer se geni v predelih, ki so bogati s ponovitvami, razvijajo precej hitreje kot preostali genom in tako predstavljajo osnovo za adaptivno evolucijo (Dong in sod., 2015). Primerjalne analize genomov so prikazale še njihovo nagnjenost k značilni arhitekturi z evolucijskim trendom k velikim genomom, ki so polni ponovitvenih zaporedij. Izkazalo se je tudi, da ti genomi vsebujejo nabore genov, ki kodirajo sekretorne proteine – v velikih primerih hkrati tudi efektorske proteine – ki delujejo kot virulenčni faktorji z modulacijo celičnih procesov v rastlini (Dong in sod., 2015).

2.3 MITOHONDRIJSKI GENOMI GLIV

Glivni mitohondrijski genomi se zelo razlikujejo po velikosti, številu genov in genomski sestavi. Npr. mitohondrij kvasovke *S. pombe* je sestavljen iz samo 19.431 baz z 10 protein-kodirajočimi geni, medtem ko mitohondrij glive *P. anserina* vsebuje 50 protein-kodirajočih genov in je velik okoli 100 kbp. Število genov v genomu pogosto ni sorazmerno z velikostjo mitohondrijskega genoma, kot npr. mitohondrij *M. perniciosa,* ki pri velikosti 109 kb vsebuje le 14 protein-kodirajočih genov. Te vrednosti so povečini

v skladu z ostalimi evkariontskimi mitohondrijskimi genomi, kjer je število možnih protein-kodirajočih genov med 3 in 67 (Haas in sod., 2011).

Poleg razlik v vsebini se mitohondrijski genomi lahko razlikujejo tudi po arhitekturi. Večino glivnih mitohondrijskih genomov tvori posamezen krog, vendar obstajajo izjeme, ki imajo mitohondrijski genom sestavljen tudi iz večih krogov, kot npr. *Spizellomyces punctatus* mitohondrijski genom, sestavljen iz treh krožnih kromosomov (58,8, 1,4 in 1,1 kb), ki vsebujejo 31, 1, in 0 protein-kodirajočih genov (Burger in sod., 2003). Posebni primeri so tudi mitohondrijskimi genomi gliv rodu *Candida*, ki so po obliki linearni genomi s telomerami in so včasih tudi sestavljeni iz več kromosomov (Kosa in sod., 2006).

Glivni mitohondrijski genomi askomicet ponavadi vsebujejo standardni nabor 14-ih ali 15-ih ohranjenih protein-kodirajočih genov, ki se povečini prepisujejo z iste DNA verige. Med temi so podenote ATP sintaze (*atp6, atp8, atp9*), apocitokrom b (*cob*), podenote citokrom oksidaze (*cox1, cox2, cox3*), podenote NADH dehidrogenaze (*nad1, nad2, nad3, nad4, nad4L, nad5* and *nad6*) in *rps* 5 ribosomalni protein. Zaradi teh lastnosti se lahko relativno preprosto primerja gensko vsebnost in zaporedje genov med mitohondrijskimi genomi iz različnih glivnih vrst skupine askomicet. Velika mera ohranjenosti pa pomaga tudi pri pripisu značilnosti mitohondrijskih genomov novih glivnih vrst znotraj te skupine (Haas in sod., 2011).

Sam pripis genomskih značilnosti predstavlja unikaten izziv, ker imajo mitohondrijski genomi drugačno sestavo od jedrnih genomov in uporabljajo drugačne zapise za prevajanje kodonov. Npr. jedrni stop-kodon UGA v mitohondrijskem genomu predstavlja aminokislino triptofan. Mitohondrijski genomi lahko uporabljajo tudi alternativne start kodone, kot so npr. AUU, AUA in UUA. Glivni mitohondrijski genomi vsebujejo tudi ne-kodirajoče RNA gene, kot npr. veliko in malo podenoto za rRNA. Ti geni so precej krajši od sorodnih jedrnih genov in jih je zato dostikrat težko identificirati. Druga oblika ne-kodirajočih RNA genov v mitohondrijskih genomih so tRNA kodirajoči geni, kateri pri glivah povečini nastopajo v naboru vseh 20 aminokislin. Najpogosteje se tRNA geni nahajajo razporejeni po genomu v gručah in dostikrat tudi v okolici rRNA genov. Zaporedje tRNA genov je prav tako povečini dobro ohranjeno med bližnje sorodnimi vrstami. Nekateri mitohondrijski genomi, ki ne vsebujejo vseh tRNA, nadoknadijo manjkajoče z uvozom iz jedrnega genoma (Haas in sod., 2002).

2.4 PREDHODNE GENOMSKE RAZISKAVE VRST *Verticillium*

Odkritja genomskih značilnosti fitopatogenih gliv se beležijo v podatkovnih bazah, ki so večinoma dostopne preko spleta, kot npr. GenBank (Benson in sod., 2005). Te baze pogosto vsebujejo redundantne nabore podatkov, ker zajemajo vire iz drugih baz, lahko pa tudi vsebujejo neskladja med podatkovnimi viri ali pa različne formate podatkov, kar sta dva izmed splošnih največjih problemov pri delu s podatki v bioinformatiki.

Za glive družine *Verticillium* je spodaj predstavljen kratek pregled trenutnih (December, 2016) stanj genomskih sestavkov, ki so dostopni v repozitorijih Genbank ter ENA.

Verticillium dahliae - 33,9 Mbp ... dokončana genom in mitohondrij po kriterijih NCBI (27.184 bp)

Verticillium dahliae JR 2- 36,2 Mbp ... ogrodje (*angl. scaffold*)

Verticillium tricorpus - 36,06 Mbp ... ogrodje

Verticillium alfalfae - 32,86 Mbp ... ogrodje

Povodi analiz teh genomov so bili različni, vendar glavni razlog v vseh primerih je bilo preučevanje dejavnikov za njihovo specifično patogenost. Nekaj razlogov in ugotovitev z različnih vidikov raziskav na teh genomih je predstavljenih spodaj:

- *V. dahliae* VdLs.17 [GCF_000150675.1]

- *V. alfalfae VaMs.102* [GCA_000150825.1]

Poleg sestave genomov in raziskave njihovih filogenetskih odnosov je bil poseben poudarek pri teh genomih posvečen preučevanju njihovih efektorjev in sekretornih proteinov. Še posebej izstopata družini sekretornih proteinov in karbohidrat-aktivnih encimov, ki sta precej razširjeni v genomih *V. dahliae* in *V. alfalfae*. Za ta genoma je značilna tudi visoka stopnja podobnosti in splošna razširjenost genomskih otočkov bogatih s ponovitvami (Klosterman in sod., 2011).

- *V. dahliae* JR2 [GCA_000400815.2]

Raziskovalni fokus tega genoma je bil usmerjen v raziskavo kromosomskih prerazporeditev, zaradi katerih nastajajo dinamične vrstno-specifične genomske regije. Te zaradi arhitekturne nestabilnosti delujejo kot vir genetske variabilnosti za vzdrževanje oz. povečevanje virulentnosti patogena. Te regije se pojavljajo na koncih kromosomov in so obogatene z retrotranspozoni, ponovitvenimi zaporedji in efektorskimi geni ter zaradi pogostih kromosomskih prerazporeditev omogočajo hiter razvoj novih efektorskih genov (de Jonge in sod., 2013).

- *V. tricorpus* [GCA_000732205.1]

Cilj raziskav tega genoma je bila predstavitev uporabe hibridnih tehnologij sekvenciranja druge in tretje generacije pri sestavi genoma ter primerjava le-tega z genomom *V. dahliae* za določitev homolognih genomskih otočkov z efektorskimi proteini. Z dodatkom tehnologij sekvenciranja tretje generacije (konkretno SMRT tehnologije podjetja Pacific Biosciences) se je proces sestavljanja genoma precej skrajšal in poenostavil, zaradi kombinacije z NGS tehnologijami pa je na koncu vseboval tudi manj napak. S tem so pokazali, da je uporaba tehnologij tretje generacije že dovolj dozorela za praktične namene, vendar je njena glavna slabost še vedno cena in relativno težka dostopnost (Seidl in sod., 2015).

2.5. POSTOPKI IN PRINCIPI V GENOMSKIH RAZISKAVAH

V sledečih poglavjih so predstavljena področja, ki se uporabljajo pri raziskovanju genomov. Sestavljanje zaporedij (*angl. assembly*), pripis genomskih značilnosti (*angl. annotation*) in filogenetske analize so osnovna področja, ki jih uporabljamo v genomskih raziskavah in so v naslednjih poglavjih tudi podrobneje predstavljena. Prav tako bo omenjeno tudi področje strojnega učenja in njegov doprinos k tehnikam, ki se uporabljajo v bioinformatiki. Znotraj področja filogenetskih analiz pa bosta predstavljena tudi evolucijska analiza, Ka/Ks koeficient ter njun pomen pri raziskovanju vplivov mutacij na genome.

2.5.1 Sestavljanje zaporedij (assembly)

Pomembna odločitev pred začetkom sestavljanja je izbira ustrezne programske opreme – prosto dostopne (večino programske opreme razvite za področje bioinformatike) ali pa komercialne programske opreme (npr. CLC Genomics Workbench). Komercialna programska oprema je ponavadi uporabniku bolj prijazna kot prosto dostopni programi in je zato tudi bolj primerna za raziskovalce z omejenimi veščinami v bioinformatiki. Slaba stran te programske opreme, razen (pogosto visoke) cene za nakup in licenciranje je, da deluje kot 'black box' rešitev, pri čemer je večinoma nemogoče dobiti vpogled v njeno delovanje ali uporabniškega nastavljanja podrobnosti delovanja algoritmov. Nekatere pogosto uporabljene programske rešitve so na voljo le ob nakupu skupaj s sekvenatorjem in so brez tega nakupa tako dostopne le preko sekvenatorskih centrov (Ekblom in Wolf, 2014).

Trenutno večina genomskih projektov uporablja WGS (*angl. whole-genome shotgun*) pristop kot strategijo za sekvenciranje genomov. V prvem koraku tega pristopa se genomska DNA razreže na kratke naključne fragmente - odčitke, kateri se potem na

podlagi uporabljene tehnologije posekvencirajo na določeno dolžino. Te odčitke se uporabi v postopku sestavljanja, pri čemer se z računskimi pristopi poskuša zaporedja ponovno sestaviti nazaj v izvorne molekule - kromosome (Miller in sod., 2010). Za pravilno sestavljanje je pomembno, da je na voljo dovolj prekrivanja odčitkov na vsaki poziciji genoma, kar imenujemo tudi visoka pokritost sekvenciranja (Ekblom in Wolf, 2014).

Za sestavljanje genoma obstajata 2 pristopa: primerjalni in *de-novo* pristop. Med prvim, ki se imenuje tudi sestavljanje s pomočjo reference, se uporabi referenčni genom iz istega organizma ali bližje sorodne vrste kot matrica za vodenje procesa sestavljanja s poravnavo odčitkov (El-Metwally in sod., 2013). Med *de-novo* sestavljanjem ni na voljo vnaprej podanih informacij o zaporedjih, kot so npr. genomske regije, transkripti in proteini in zatorej ta pristop pomeni najbolj strogo obliko sestavljanja ter se uporablja pri sestavljanju genomov, kateri nimajo na voljo sorodnih posekvenciranih genomov (Miller in sod., 2010).

Trenutno uporabljene tehnologije za NGS sekvenciranje DNA z WGS pristopom imajo skupno omejitev, ki predstavlja velik izziv - dolžine odčitkov so dosti krajše tudi od najmanjših genomov. Ta pomanjkljivost se poskuša premostiti z večkratnim vzorčenjem tarčnega genoma iz naključnih pozicij (Miller in sod., 2010). Težavnost izziva je odvisna tudi od tehnologije sekvenciranja, ker je delež ponavljajočih odčitkov odvisen od njihove dolžine. V eni skrajnosti, če bi bili odčitki dolgi samo eno bazo, bi bili vsi odčitki ponavljajoči; v drugi skrajnosti, če bi lahko prebrali celoten kromosom naenkrat, potem ponavljajoče regije ne bi predstavljale problema. Med temi skrajnostmi se delež unikatnih zaporedij povečuje z dolžino odčitkov, dokler vsako zaporedje v genomu ni unikatno. Vendar realni genomi vsebujejo zapletene strukture ponovitev, zaradi česar je nekatera zaporedja skoraj nemogoče pravilno sestaviti (Schatz in sod., 2010).

Razrešitev mej med daljšimi regijami ponovitev je včasih mogoča z uporabo posameznih odčitkov, ki premoščajo instance ponovitev z zadostnim številom unikatnih zaporedij na obeh straneh ponovljive regije »oz. premostnikov« (primer takšnih so denimo dolgi odčitki Oxford Nanopore in Pacific Biosciences) ali pa z uporabo knjižnic parnih odčitkov (tradicionalno v uporabi za združevanje v ogrodja pri NGS tehnologijah). Pri slednjih sta za popolno razrešitev ponovitvenih regij ponavadi potrebna 2 tipa odčitkov: pari, ki premoščajo ponovljive regije z unikatnimi zaporedji na obeh straneh in pari, ki imajo samo en konec v ponovljivi regiji (Miller in sod., 2010).

2.5.1.1 Algoritmi grafov za sestavljanje zaporedij

Graf je matematična abstrakcija, ki se pogosto uporablja na področju računalniških znanosti in se tipično predstavi kot nabor točk ter vezi med njimi. Te točke in vezi se drugače poimenujejo tudi vozlišča in povezave. Če povezave med vozlišči potekajo samo v eno smer, potem se graf imenuje usmerjeni graf. Vsaka usmerjena povezava predstavlja povezavo iz enega izvornega vozlišča v eno ponorno vozlišče. Zbirke povezav tvorijo poti, ki obiščejo vozlišča v nekem zaporedju, tako da ponorno vozlišče ene povezave tvori izvorno vozlišče za sledeča vozlišča. Poseben tip poti se imenuje preprosta pot, katera vsebuje samo unikatna vozlišča (vsako vozlišče se pojavi samo enkrat). Preprosta pot po definiciji ne more sekati samo sebe in dodatno se lahko zahteva tudi, da je ne sme sekati nobena druga pot. Vozliščem in povezavam se lahko pripiše različne atribute in semantike, kar je predvsem uporabno pri praktičnih aplikacijah grafov. V kontekstu grafov sestavljanje genoma predstavlja problem redukcije grafa. Za večino optimalnih redukcij grafov ne poznamo učinkovitih rešitev in se zaradi tega programi za sestavljanje zanašajo na hevristične algoritme in približke, da odstranijo redundanco, popravijo napake, zmanjšajo kompleksnost, povečajo preproste poti in čim bolj poenostavijo graf (Miller in sod., 2010).

Graf prekrivanj pri procesu sestavljanja genoma predstavlja odčitke sekvenciranja in njihova prekrivanja. Ta morajo biti vnaprej preračunana preko (računsko zahtevnih) parnih poravnav zaporedij. Konceptualno ima graf vozlišča, ki predstavljajo odčitke ter povezave, ki predstavljajo prekrivanja. V praksi ima lahko graf dodane še posebne elemente ali atribute, ki razločijo 5' in 3' konce odčitkov, zaporedja odčitkov in njihove reverzne komplemente, dolžine odčitkov, dolžine prekrivanj in tipe prekrivanj (predpona, pripona, popolno). Povezave skozi graf so potencialne soseske, ki se jih lahko pretvori v zaporedja in imajo lahko tudi zrcalne poti, ki predstavljajo reverzne komplemente zaporedij (Miller in sod., 2010).

Programe za sestavljanje zaporedij po WGS pristopu poznamo v več kategorijah, od katerih so vse osnovane na grafih:

- Overlap/Layout/Consensus (OLC) metode, ki temeljijo na grafih prekrivanj;
- De Bruijn Graph (DBG) metode, ki so osnovane na grafih K-mer;
- Greedy graph metode (GG), ki temeljijo na mešanici OLC in DBG pristopov.

2.5.1.1.1 Overlap/Layout/Consensus pristop

OLC je bil najbolj tipičen pristop za programe ob nastopu Sangerjeve tehnologije sekvenciranja in je optimiziran za velike genome. Pri njem se uporabijo poravnave med odčitki tipa vsi-proti-vsem, da se pripravi graf povezav odčitkov in na podlagi poti med njimi se nato sestavijo soseske. Zaradi ponovljenih odčitkov, napak in ponovitev se lahko v grafu pojavi več poti skozi povezave in na podlagi razlik v načinu razrešitve teh poti, se je razvilo več različnih metod oz. implementacij tega pristopa (MacLean in sod., 2009).

OLC programi za sestavljanje uporabljajo graf prekrivanj in za svoje delovanje izvajajo 3 korake:

1) Odkritje prekrivanj z vsak-proti-vsem parnimi poravnavami odčitkov
Za boljšo učinkovitost poravnav se uporablja seed & extend hevristični algoritem. Pri tem koraku se vnaprej izračuna vsebnost K-mer v vseh odčitkih, izbere kandidate z dobrim prekrivanjem na podlagi skupnih K-mer in izračuna poravnave z uporabo K-mer kot semen za poravnavo. Odkritje prekrivanj je občutljivo na velikost K-mer, najmanjšo dolžino prekrivanja in najmanjšo zahtevano stopnjo identičnosti za prekrivanje. Ti trije parametri najbolj vplivajo na robustnost, upoštevajo pa se tudi napake v določevanju baz in sekvenciranju z nizko pokritostjo.

2) Izgradnja in poenostavitev grafa prekrivanj
Ta korak pripravi približke postavitev odčitkov v soseske. Graf prekrivanj ne potrebuje informacij o določevanju baz pri sekvenciranju, zaradi česar se lahko grafi velikih genomov vzdržujejo tudi v praktičnih količinah računalniškega pomnilnika.

3) Večkratne poravnave za določitev natančnih postavitev in skupnega zaporedja
Trenutno ni poznana nobena učinkovita metoda za izračun optimalne poravnave več zaporedij, zato določitev skupnega zaporedja uporablja progresivne parne poravnave, ki jih usmerja približna postavitev odčitkov. Ta faza mora shraniti podatke v pomnilniku in lahko poteka vzporedno, ločeno po soseskah (Miller in sod., 2010).

Spodnji primer predstavlja vizualno ponazoritev grafa prekrivanj.

Izvorno zaporedje

```
GGTACTGCCGGACTGAATACTGGT
```

Odčitki

```
GGTAC  CCGGACTGAAT CTGGT
  TACTG   GACTGAA ACTG
 GTAC GCCGG   GAATACT
GGTACTGCC  ACTGAATAC GGT
```

Sestava skupnega zaporedja po OLC pristopu iz vhodnih odčitkov:

```
GGTACTGCC
GGTAC
 GTAC
  TACTG
      GCCGG
       CCGGACTGAAT
          GACTGAA
           ACTGAATAC
              GAATACT
                  ACTG
                   CTGGT
                     GGT
GGTACTGCCGGACTGAATACTGGT
```

Končno zaporedje je rezultat parnih poravnav vseh odčitkov in njihovega iterativnega združevanja. Iz primera je tudi razvidno, da se daljše odčitke precej lažje umesti v končno zaporedje, kot pa krajše.

2.5.1.1.2 de Bruijn Graph pristop

Najbolj razširjen pristop za sestavljanje genomov po WGS pristopu za NGS podatke je uporaba de Bruijn-ovih grafov. Ti temeljijo na grafih K-mer in se s pridom uporabljajo pri velikem številu kratkih odčitkov, ki jih generirajo NGS tehnologije. Graf K-mer, v nasprotju z grafom prekrivanj, ne potrebuje vsako-proti-vsem parnih poravnav za določanje prekrivanj, ne potrebuje posameznih odčitkov ali njihovih prekrivanj in lahko učinkovito upravlja z redundantnimi zaporedji. (Miller in sod., 2010).

Pri de Bruijn-ovem grafu vozlišča predstavljajo vsa podzaporedja določene dolžine (K-mere), ki izhajajo iz originalnega zaporedja, povezave med njimi pa vsa prekrivanja določene dolžine med zaporednimi podzaporedji. Možna je tudi obratna formulacija, kjer vozlišča predstavljajo prekrivanja in povezave predstavljajo K-mere. Pri aplikaciji na WGS sestavljanje zaporedij, ta graf predstavlja vhodne odčitke. Vsak odčitek povzroči nastanek nove poti v grafu, pri čemer odčitki s popolnim prekrivanjem povzročijo nastanek skupne poti. Pozitivna posledica tega je, da se popolna prekrivanja odkrijejo implicitno brez računanja parnih poravnav med odčitki. V primerjavi z grafi prekrivanja so grafi K-mer bolj dovzetni za napake pri sekvenciranju in ponovitve, katere inducirajo cikle v grafu, zaradi tega povzročijo več možnih načinov njegove rekonstrukcije (Pevzner in sod., 2001).

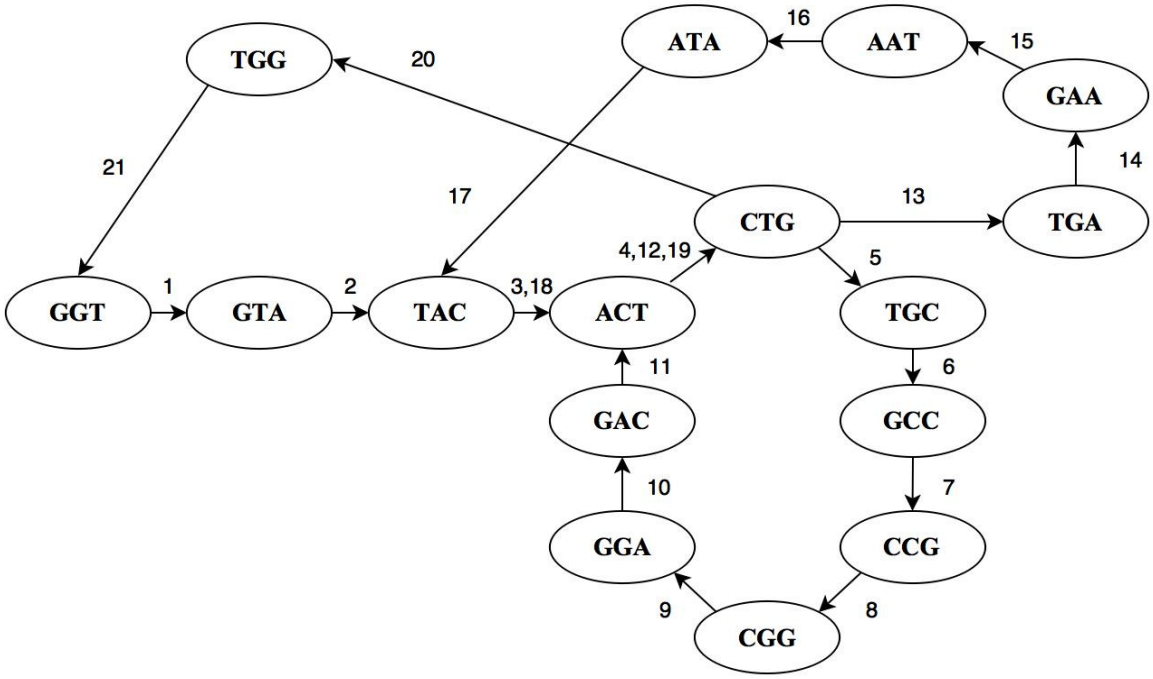
Spodnji primer predstavlja vizualno ponazoritev preprostega grafa K-mer in ciklov, ki nastanejo zaradi ponovljivih zaporedij.

Izvorno zaporedje

```
GGTACTGCCGGACTGAATACTGGT
```

Odčitki

```
GGTAC  CCGGACTGAAT CTGGT
  TACTG   GACTGAA ACTG
 GTAC GCCGG   GAATACT
GGTACTGCC  ACTGAATAC GGT
```



Slika 1: Sestava skupnega zaporedja z de Bruijn graf pristopom in K-merami dolžine 3
Figure 1: Consensus sequence assembly with the de Bruijn graph method and K-mers of length 3

Sam računski proces sestavljanja genoma po WGS pristopu z uporabo grafov K-mer združuje 4 zaporedne podprocese:

1) predprocesno filtriranje;
2) proces izgradnje grafa;
3) proces poenostavitve grafa;
4) postprocesno filtriranje.

Korak predprocesnega filtriranja je odgovoren za odkritje in morebitno popravo odčitkov z napakami, preden se začne proces sestavljanja. Proces izgradnje grafa je odgovoren za izdelavo modela, ki se uporabi za organizacijo zaporedij odčitkov v kompaktno obliko in izdelavo daljših odčitkov med sestavljanjem. Proces poenostavitve grafa se uporabi, da poenostavi model z zmanjšanjem števila vozlišč in povezav v grafu ter da odstrani napačne povezave. Korak postprocesnega filtriranja nato sestavi soseske, odkrije in odstrani napačno sestavljene soseske ter pravilne sestavke združi v ogrodja. Pri tej zadnji stopnji se morebitni parni odčitki vgradijo v soseske, tako da naredijo graf povezanosti sosesk ali pa uporabijo že predhodno sestavljenega (v drugi stopnji) s posodobljenimi informacijami. Z vključitvijo parnih odčitkov se tako lahko odkrijejo napačno sestavljene soseske in nepravilne ponovitve. Trenutni programi za sestavljanje

genomov po WGS pristopu imajo vse ali pa samo nekatere od teh stopenj. Npr. pomanjkanje korakov lastnega pred- in post- procesiranja nadomestijo z uporabo samostojnih orodij za predprocesno filtriranje (orodja za popravo napak) in za postprocesno filtriranje (orodja za sestavo ogrodij). Nekateri programi tudi preložijo korak odpravljanja napak do stopnje poenostavitve grafa, ker nekatere napake niso opazne vse dokler se graf ne začne graditi, npr. razlikovanje polimorfizmov od napak pri sekvenciranju (El-Metwally in sod., 2013).

Predprocesno filtriranje

Pred sestavljanjem zaporedja je priporočljivo, da se oceni določene parametre odčitkov (kot npr. celotna GC vsebnost, kvalitete klicev baznih parov, pogostost ponovitev, odstotek podvojenih odčitkov, ...). Za ta proces so na voljo orodja, ki podajo informativne statistike, na podlagi katerih se potem naredijo ustrezne odločitve za predprocesiranje. Kot alternativa tem orodjem je možno tudi samostojno popravljanje na podlagi frekvence K-mer, kadar je to ustrezen pristop za dani nabor podatkov. Optimalna strategija za predprocesiranje odčitkov pri vseh projektih ne obstaja, ker je potreba po filtriranju podatkov odvisna od posameznega projekta in uporabljenega programskega cevovoda za sestavljanje genoma. Nekateri programi potrebujejo kot vhodne podatke samo surove odčitke brez filtriranja po kakovosti, ker je popravljanje njihovih napak že implementirano znotraj vnaprej pripravljenega programskega cevovoda (El-Metwally in sod., 2013).

Zaporedja začetnih oligonukleotidov, adapterjev in vektorjev iz priprave knjižnic so zelo pogosto prisotna v podatkih (tudi če sekvenčni center trdi, da jih je odstranil) in jih je priporočljivo še enkrat preveriti pred začetkom dela. Pri sekvenciranju z Illumina tehnologijo, se doda DNA PhiX faga v reakcijo za sekvenciranje, za kalibracijo ocene kakovosti zaporedij. Če se ta vir prekomerno kontaminiranih zaporedij ne odstrani, se lahko proces sestavljanja oslabi (zaradi velikega števila odčitkov kontaminantov v primerjavi s preučevanim genomom) in zaradi tega nastanejo himerne ter kontaminirane soseske. Najlažji način odstranitve znanih vektorjev iz surovih podatkov je z uporabo programov za kartiranje kratkih odčitkov in izbris vseh fragmentov, ki se kartirajo na kontaminirana zaporedja (Ekblom in Wolf, 2014).

Sestavljanje genoma

Orodja za sestavljanje genomov se precej razlikujejo v svojih učinkovitostih glede na merila hitrosti, skalabilnosti in kvaliteti končnih genomskih zaporedij. Čeprav nekatere metode sestavljanja odločno prekosijo druge, je vseeno težko ugotoviti, katero orodje bi lahko bilo najprimernejše za določen nabor podatkov. Vsak projekt sestavljanja genoma

je unikaten z ozirom na generirane podatke in tudi na npr. velikost, bazno sestavo, vsebnost ponovitev in stopnjo polimorfizmov v genomu. Na voljo so številni programi za *de-novo* sestavljanje WGS genomskih podatkov in stalno prihaja do razvoja novih, pri čemer se nekateri algoritmi osredotočajo na minimiziranje napačnih sestavkov, medtem ko drugi želijo maksimizirati povezljivost sosesk (pogosto tudi za ceno kakovosti). Večina algoritmov za sestavljanje deluje optimalno z določeno porazdelitvijo velikosti knjižnic, tako da se je pomembno odločiti za strategijo sestavljanja že med načrtovanjem projekta in/ali tekom sekvenciranja. Informacije o tem se lahko pridobi iz primarne literature in internetnih strani, namenjenih programom za sestavljanje genomov, prav tako pa tudi z različnih spletnih forumov, na katerih so na voljo razprave o najsodobnejših metodah in so podane tudi praktične izkušnje drugih raziskovalcev o njihovi uporabi (Ekblom in Wolf, 2014).

Pri izbiri programske opreme je pomemben dejavnik tudi količina sekvenčnih podatkov in razpoložljiva računalniška oprema. Metode, osnovane na de Bruijn-ovih grafih, v splošnem potrebujejo velike količine računskega pomnilnika (RAM-a). Odvisno od količine sekvenčnih podatkov, lahko sestavljanje genomov v velikostnem rangu sesalcev (3 Gbp) zasede več terabajtov internega pomnilnika. Če ni na voljo dovolj zmogljive računalniške opreme, so na voljo alternative v obliki skupnih nakupov opreme, skupnih projektov z bioinformatskimi skupinami, uporaba storitev (SaaS, IaaS, ...) splošnega komercialno dostopnega računalništva v oblaku (npr. Google Cloud, Amazon EC2) ali pa namenskih oblačnih storitev za bioinfomatiko (iPlant Collaborative, Galaxy Cloud, DNAStar Cloud). V Sloveniji imamo za ta namen možnost koristiti tudi oblačno storitev Akademske in raziskovalne mreže Slovenije (Arnes) s strežnikom po meri.

Večino programov za sestavljanje genomov poleg izgrajenega grafa izvozi tudi nabor nesestavljenih ali delno sestavljenih sosesk. Najbolj pogosto uporabljen podatkovni format je FASTA, kjer so skupna zaporedja sosesk predstavljena kot nizi znakov A, C, G, T in včasih tudi drugih znakov s posebnimi pomeni. Npr. pomišljaj lahko predstavlja dodatne baze, ki so bile izpuščene iz skupnega zaporedja, ampak so prisotne v manjšini uporabljenih odčitkov. Najbolj pogosti izmed teh znakov so nizi N-jev v vrzelih med soseskami v ogrodjih, ki označujejo predvideno dolžino vrzeli med soseskami (Miller in sod., 2010).

Za združevanje sosesk v ogrodja (*angl. scaffolding*), se uporabijo knjižnice dolgih DNA fragmentov, ki lahko premoščajo več kilobaz genomskega zaporedja. Glede na tehnologijo in specifike priprave knjižnic, se te poimenujejo kot knjižnice parnih-koncev, sosednjih-parov ali preskočne knjižnice. Če se konci zaporedij več neodvisnih

fragmentov nahajajo na dveh različnih soseskah, se te soseske združijo v ogrodje. Pričakovana dolžina fragmentov iz uporabljene knjižnice oceni informacijo o fizični razdalji med dvema soseskama in vrzel med njima se napolni z znakom 'N'. Kasnejše metode za zapiranje vrzeli lahko zapolnijo informacije o manjkajočih bazah, v idealnem primeru z uporabo dolgih odčitkov, ki segajo preko ponavljajočih regij. V zadnjem koraku se ogrodja lahko povežejo v povezane skupine ali pa kartirajo na kromosome (Ekblom in Wolf, 2014).

Postprocesno filtriranje in kontrola kakovosti

Uspešno zaključeno sestavljanje genoma je nato podvrženo oceni rezultata s koraki kontrole kakovosti in na podlagi njenih ugotovitev se potem osnuje najprimernejša strategija za nadaljnji postopek postprocesnega filtriranja. V najširšem pomenu se lahko postopki kontrole kakovosti razdelijo v pristope, ki potrebujejo dodatne informacije iz zunanjih podatkov in tiste, ki delujejo samo na podlagi lastnih podatkov sestavka. En izmed najbolj kakovostnih virov zunanjih informacij za ta korak je soroden referenčni genom (Clark in sod., 2013).

Kadar referenčni genom ni na voljo, potem se lahko kakovost sestavljenega zaporedja oceni s poravnavno surovih odčitkov, pri čemer pridobimo informacije o globini pokritosti ter konsistenci parnih odčitkov. S temi informacijami se lahko zazna nekatere napačne sestavke kot npr. razlike v posameznih bazah, zlitja ali razširitve ponovitvenih zaporedij in podvojitve segmentov. Nekatera orodja lahko te razlike še dodatno statistično ovrednotijo s sintezo večih parametrov, kot so vhodni odčitki, njihove k-merne kompozicije, globine pokritosti zaporedja, dolžine insertov in kakovosti kartiranj posameznih baz. Z integracijo teh podatkov se posledično lahko odkrijejo tudi bolj specifične napake, kot so npr. indeli ali pa napake v številu kopij ter himerni metagenomski sestavki. Ena izmed najbolj pogosto predstavljenih ocen kakovosti sestavljenega genoma so porazdelitve sosesk, ki se lahko ocenijo z referenčnim genomom ali brez njega. Pri njih se ocenijo parametri kot so: št. sosesk, največja soseska, celotna dolžina sestavka, Nx ($0 < x < 100$; najpogostejša vrednost je N50), število neporavnanih sosesk, število sosesk, ki se kartira na več lokacij po referenčnem genomu ter drugi. Pri sintezi teh podatkov z drugimi viri se lahko dodatno tudi statistično oceni verjetnost pravilne ali pa napačne sestave izbrane soseske (Clark in sod., 2013).

Kadar je bližnje soroden referenčni genom na voljo, potem lahko naredimo tudi bolj specifične ocene novo sestavljenega genoma. Če se sestavek dobro ujema z referenco, potem je možno ovrednotiti tudi potencialne napačne sestavke in strukturne variacije, ki

so lahko posledica napak pri sekvenciranju, lahko pa tudi dejanskih strukturnih variacij, kot npr. prerazporeditev, večjih indelov ali pa variabilnega števila ponovitvenih zaporedij (Gurevich in sod., 2013).

Na podlagi referenčnega genoma lahko pridobimo tudi precej informativne ocene o vsebnosti genoma, kot npr. (Gurevich in sod., 2013):

- delež poravnanega sestavljenega genoma na referenčno zaporedje;
- razmerje števila poravnanih baz sestavka v primerjavi s številom poravnanih baz na referenci. Če delež precej odstopa od povprečja 1 navzgor, potem to nakazuje na podvojitve v sestavku;
- odstotek vsebnosti GC;
- št. neujemanj na dolžino (npr. 100 kb) poravnanega zaporedja;
- št. delno in celotno sestavljenih genov glede na referenčno zaporedje.
  Lahko pa tudi ocenimo, v koliki meri (delno ali v celoti) so se sestavili funkcionalni genomski elementi, kot so geni in operoni.

Zelo dober pristop za izboljšanje genomskih sestavkov ali pa tudi za vodenje sestavljanja genomov visoke kakovosti je uporaba genomskih kart. Te lahko na podlagi razlik s sestavki pomagajo pri odkritju napačno sestavljenih regij in podajo neodvisno oceno o kakovosti sestavkov, pri čemer so na voljo v več oblikah, kot npr. fizične karte, optične karte ter genetske karte (Madoui in sod., 2016).

2.5.1.1.3 Greedy Graph pristop

Požrešni (Greedy) pristop aplicira na vhodne podatke samo eno osnovno operacijo: na podani odčitek ali soseko doda še en odčitek ali sosesko z najboljšim prekrivanjem. Osnovna operacija se tako ponavlja, dokler je ni več možno izvesti, pri čemer vsaka operacija uporabi trenutno najboljše možno prekrivanje za odločitev pri združevanju. Na podlagi tega soseske rastejo s "požrešnim" povečevanjem, tako da se zmeraj združijo z zaporedjem, ki ga ocenitvena funkcija oceni za najboljšega. To je hkrati tudi slabost požrešnih algoritmov, saj se ti lahko ujamejo v lokalnem maksimumu, če se trenutna soseska poveča na podlagi odčitkov, ki bi druge soseke povečali še bolj. Požrešni algoritmi lahko za optimizacijo svojega delovanja ustvarjajo instanco samo enega prekrivanja vsakega odčitka, ki ga preverijo, lahko pa tudi zavržejo vsako prekrivanje takoj po podaljšanju soseske (Miller in sod., 2010).

Kot tudi drugi programi za sestavljanje zaporedij, potrebujejo požrešni algoritmi mehanizme, da se izognejo vključevanju lažno-pozitivnih prekrivanj v soseske. Prekrivanja, ki jih inducirajo ponovitve, imajo lahko višje točkovanje kot prekrivanja, ki

jih inducirajo skupna mesta izvora na zaporedju, zaradi česar bo program, ki gradi na podlagi lažno-pozitivnih prekrivanj, združil nesorodna zaporedja s ponovitvami in tako naredil himerna zaporedja (Miller in sod., 2010).

2.5.1.2 Sestavljanje organelnih genomov

Podobni postopki, kot so bili predstavljeni v poglavju za WGS sestavljanje jedrnih genomov, veljajo tudi za WGS sestavljanje mitohondrijskih genomov. Dodaten korak pri slednjih je ta, da jih je potrebno ločiti od zaporedij jedrnega genoma, kar se lahko naredi ob izolaciji DNA z obogatitvijo mitohondrijske DNA v vzorcu, ali pa po sestavljanju s filtriranjem ustreznih mitohondrijskih zaporedij. Proces filtriranja je mogoč zaradi večjega števila kopij mitohondrijskih genomov znotraj celice in posledične večje pokritosti mitohondrijskih sosesk v primerjavi z jedrnimi pri sestavljanju. Pri postopku s filtriranjem po sestavljanju, se najprej razvrstijo nastale soseke glede na pokritost in se jih tako loči po njihovi pogostnosti v celici (cpDNA ~ 2,000 – 3,000x, mtDNA ~200x, ~20x jedrna DNA). Te se nato lahko identificira glede na podobnost s sorodnimi organelnimi genomi z algoritmom BLAST (Camacho in sod., 2009).

Primer takšnega procesa ponuja cevovod Iterative Organelle Genome Assembly (IOGA), ki sestavi parne odčitke v nabor kandidatnih sestavkov, izmed katerih nato z ocenitvijo verjetnosti izbere najustreznejšega (Bakker in sod., 2016).
Proces poteka v več korakih (Bakker in sod., 2016):

1) Popravljanje regij nizke kakovosti na odčitkih in odstranjevanje morebitnih adapterjev;
2) Filtriranje odčitkov, ki izhajajo iz organelnih genomov, preko kartiranja na zbirko referenčnih organelnih genomskih zaporedij;
3) Sestavljanje organelnega genoma iz filtriranih odčitkov na podlagi K-mer velikosti 37-97;
4) Najustreznejši sestavki se izberejo na podlagi N50 statistike in se nato uporabijo kot nova referenca za iskanje tarčno-specifičnih odčitkov, ki niso bili izbrani v prvi iteraciji;
5) Prejšnji korak se ponavlja, dokler se ne doda več novih odčitkov k referenci;
6) Sestavljanje organelnega genoma na podlagi končne zbirke odčitkov, z uporabo podobnega obsega velikosti K-mer kot v koraku 3;
7) Ocena sestavkov iz korakov (3) in (5) z Assembly Likelihood Estimation (ALE). Tisti, ki je najbolje ocenjen, se izbere kot končni sestavek.

Navkljub veliki pokritosti, ki jo je mogoče doseči s sodobnimi tehnologijami za sekvenciranje, lahko dodaten problem predstavlja nastop heteroplazmije v celici. Zaznava le-te v mitohondrijih je še vedno izziv, ker je pogosto težko ločiti med resničnimi alelnimi mesti in potencialnimi napakami sekvenciranja. Trenutne metodologije, ki se spopadajo s tem problemom, temeljijo na štetju variant po kartiranju odčitkov na referenco in aplikaciji različnih mej, da se odstrani šum iz podatkov. Eno izmed možnih implementacij določitve te meje ponuja delo, pri katerem so uporabili simulacije sekvenciranja klonalnih vzorcev (bakteriofaga X174) in poskusili zaznati mesta heteroplazmije v umetno mešanih vzorcih. Poleg tega so uporabili tudi ocene kakovosti klicev baz in zahtevali validacijo heteroplazmij z vsaj dvema odčitkoma na vsaki verigi (Goto in sod., 2011).

2.5.2 Strojno učenje v bioinformatiki

Metode strojnega učenja so pogosto uporabljene na področju bioinformatike, kjer so na voljo veliki nabori pogosto neurejenih in heterogenih podatkov. Implementirane so na precej raznovrstnih področjih in podobnih družinah problemov, kot obstajajo na področju podatkovnega rudarjenja (uvrščanje in razvrščanje podatkov v skupine, odkrivanje zakonitosti v podatkih, napovedovanje, ...).

Spodaj je navedenih nekaj analiz za posamezna področja bioinformatike, ki so pogosto implementirana s postopki strojnega učenja (Larran in sod., 2005):

GENOMIKA
- *prepoznavanje vzorcev* – vezavna mesta za transkripcijske faktorje, promotorska mesta, ponovitve;
- *iskanje genov* – napoved kodirajočih zaporedij, alternativnega izrezovanja, cepitvenih mest;
- *funkcije genov* – primerjava funkcij, napoved funkcij;
- *napoved strukture RNA molekul*;

TRANSKRIPTOMIKA
- *preprocesiranje in analiza podatkov* – mikromreže, NGS podatki, slike mikromrež;

PROTEOMIKA
- *napovedi struktur, funkcij in lokalizacij proteinov ter njihovih medsebojnih interakcij*;

EVOLUCIJA
- *izgradnja filogenetskih dreves*;

SISTEMSKA BIOLOGIJA
- *modeliranje omrežij* – signalnih, metabolnih, genskih;

TEKSTOVNO RUDARJENJE
- *povezovanje identifikatorjev med različnimi podatkovnimi viri*;
- *izboljšave pripisanih genomskih značilnosti na podlagi združevanja virov*;

DRUGE APLIKACIJE
- *preprocesiranje in analiza podatkov masne spektrometrije*;
- *analiza biomedicinskih slik*;
- *izdelava začetnih oligonukleotidov*;

V naslednjem poglavju sledi krajša predstavitev posameznih napovednih modelov, ki se uporabljajo na teh področjih in tudi podrobnejši vpogled v eno izmed najpogostejših metod v uporabi za pripis strukturnih značilnosti genov v genomiki.

2.5.2.1 Skriti markovski modeli

Ena izmed najbolj uporabljanih metod strojnega učenja, ki se je aplicirala na področje računske genomike, so skriti markovski modeli (HMM). Ta metoda združi multinomialne modele z modeli markovskih zaporedij in algoritme za dinamično programiranje. Njena glavna prednost je zaznavanje vzorcev, ki nimajo rigidno definiranih struktur, kakršne pogosto najdemo v bioloških zaporedjih (predvsem DNA in proteinskih zaporedjih). Metoda je uporabna za več namenov na področju računske genomike (Eddy, 2011). Spodaj so našteta najbolj pogosta (Cristianini in Hahn, 2006):

- **Segmentacija** – Genska in proteinska zaporedja lahko vsebujejo področja, ki imajo precej raznolike kemijske lastnosti. HMM pomagajo pri določitvi natančnih mej med takšnimi regijami.

- **Poravnava več zaporedij** – Poravnave več zaporedij se pogosto najlažje izračunajo z zmanjšanjem kompleksnosti vsaka-proti-vsem poravnavam v relativno preproste ena-proti-vsem poravnavam. HMM naredijo to nalogo še lažjo z definiranjem profilnih HMM, na katere se lahko poravnajo nova dodatna zaporedja. Ti profilni HMM prav tako omogočajo hitro določevanje proteinske funkcije in se lahko smatrajo kot opis poravnave več zaporedij in tudi kot model za družino zaporedij.

- **Napoved funkcije** – Pogosto navadna poravnava zaporedij ne omogoča trdnih napovedi funkcije proteinov, ker sama podobnost med zaporedji ne pomeni nujno tudi podobnosti v funkcionalnosti. HMM namesto tega omogočajo statistično ovrednoteno oceno funkcije proteinov ali pa uvrstitev proteinov v družine z neznano funkcijo. Na voljo je že veliko javnih podatkovnih baz, ki uporabljajo HMM za ta korak pripisa genomskih značilnosti.

- **Iskanje genov** – Metoda HMM je zelo uporabna pri določitvi strukture evkariontskih genov, ki nimajo rigidno določene sestave. Zaradi svoje fleksibilnosti se lahko uporabi tudi za iskanje psevdogenov, ki so skoraj identični funkcionalnim genom, le da imajo znotraj svojega zaporedja enega ali več stop kodonov, ki prekinejo običajno strukturo gena.

Definicija HMM in princip delovanja

Osnovna ideja delovanja HMM je modeliranje zaporedja, ki bi ga lahko generirala markovska veriga. Na vsaki poziciji zaporedja ima markovska veriga neznano (skrito) stanje – vse kar lahko opazimo, so le simboli, ki jih generira glede na multinomialno porazdelitev, katera je odvisna od tega stanja. Drugače povedano: informacija, ki jo dobimo o skriti markovski verigi, je posredna, z veliko verjetnostjo tudi podvržena šumu. Zaporedje, ki ga poskušamo analizirati, je zatorej modelirano kot rezultat dvojno naključnega procesa: enega, ki generira skrito markovsko verigo in enega, ki spremeni to skrito verigo v vidno zaporedje. Slednji proces sledi multinomialni porazdelitvi: v vsakem skritem stanju se uporabi drug nabor parametrov za proizvodnjo vidnega zaporedja. Ključno pri uporabi HMM-jev je, da se upošteva to zaporedje, ki vsebuje šum in se uporabi za inferenco prisotnih skritih stanj (Eddy, 2011).

HMM-ji imajo nabor dveh pomembnih parametrov: verjetnosti tranzicij in verjetnosti emisij. Parameter verjetnosti tranzicij opisuje verjetnost, s katero markovska veriga naredi prehod v katero izmed drugih skritih stanj. Ti prehodi so lahko poljubno pogosti ali redki in veriga lahko prehaja med dvema ali več skritimi stanji. Parameter verjetnosti emisij opisuje verjetnosti, s katerimi se simboli v vidnem zaporedju generirajo v vsakem izmed skritih stanj. Vsako izmed teh skritih stanj bi naj bilo sposobno generirati enake simbole, vendar v različnih frekvencah. Število emitiranih simbolov je odvisno od preučevanega zaporedja in je lahko 1 ali več (Eddy, 2011).

HMM-ji imajo parametre shranjene v dveh matrikah. Ena je matrika verjetnosti tranzicij (tranzicijska matrika) - T in druga je matrika verjetnosti emisij (emisijska matrika) – E. Tranzicijska matrika (1) ima dimenzije N x N, kjer N predstavlja število skritih stanj.

Emisijska matrika (2) ima dimenzije *N* x *M*, kjer *M* predstavlja število simbolov v vidnem zaporedju (Cristianini in Hahn, 2006). Te matrike so definirane kot:

$$T(k,l) = P(h_i = l \mid h_{i-1} = k) \quad \text{... (1)}$$

$$E(k,b) = P(s_i = b \mid h_i = k) \quad \text{... (2)}$$

Z besedami povedano: verjetnost trenutnega skritega stanja *l*, glede na to da je bila prejšnja pozicija v stanju *k*, je podana v tranzicijski matriki z vnosom *T (k,l)* in verjetnost emisije simbola *b* je določena z multinomialnim modelom, ki pripada stanju *k*, kateri je podan v emisijski matriki z vnosom *E (k,b)*. Zaporedje skritih stanj, ki jih je generiral markovski proces, imenujemo *h*; zaporedje simbolov, ki jih je generiral HMM, imenujemo *s*. Za zaporedja privzamemo dolžino *n*. Za popolnoma določen model se mora deklarirati še verjetnosti začetnih stanj markovskega procesa, ki jih označimo s *T (0, k) = P (h_1 = k)* (Cristianini in Hahn, 2006).

Izgrajen HMM se nato lahko uporabi za predvidenje zaporedja najbolj verjetnih skritih stanj, ki so generirala vidno zaporedje. Stohastični proces, za katerega privzamemo da generira zaporedje, je sledeč: skrito zaporedje je generiral markovski proces. V vsakem različnem stanju se uporabi drug multinomialen model glede na emisijske parametre, ki so povezani s skritim stanjem, da se je generiralo vidno zaporedje. To pomeni, da za vsako pozicijo skritega zaporedja, model emitira viden simbol neodvisno, iz ustrezne multinomialne porazdelitve (Cristianini in Hahn, 2006). Verjetnost podanega zaporedja se tako lahko izračuna na naslednji način:

$$P(h) = P(h_1)\prod_{i=2}^{n} P(h_i \mid h_{i-1}) = T(0,h_1)\prod_{i=2}^{n} T(h_{i-1},h_i) \quad \text{... (3)}$$

$$P(s \mid h) = \prod_{i=1}^{n} P(s_i \mid h_i) = E(h_i,s_i) \quad \text{... (4)}$$

To pomeni, da je celotna verjetnost skritega zaporedja *h* produkt posameznih verjetnosti stanj na vsaki poziciji v zaporedju (3), pri čemer verjetnosti prehodov med stanji določa tranzicijska matrika. Enako velja za verjetnost vidnega zaporedja z znanim skritim stanjem na vsaki poziciji (4), pri čemer verjetnost emisije posameznega simbola določa emisijska matrika (Cristianini in Hahn, 2006).

Kot najbolj preprost primer analize genomskega zaporedja s HMM-ji bi bil HMM z 2 skritimi stanji »kodirajoče zaporedje« in »nekodirajoče zaporedje«, vendar so zaradi

neodvisnosti od števila skritih stanj HMM-ji precej robustni in je z njimi možno modelirati precej bolj kompleksna zaporedja.

2.5.2.2 Drugi napovedni modeli

Poleg HMM se je na področje računske genomike apliciralo še vrsto drugih statističnih metod za prepoznavanje vzorcev (npr. kodirajočih zaporedij) in postopkov uvrščanja v bioloških zaporedjih:

- **Nevronske mreže** (*angl. neural networks*) – Ta metoda simulira omrežje nevronov, ki lahko preko izmenjave signalov izuči model in poda kompleksne napovedi. Nevroni si medsebojno izmenjujejo sporočila, pri čemer so povezave med njimi utežene in se lahko preko učenja modela spreminjajo ter tako sčasoma model naučijo kompleksnih pravil. Ena izmed implementacij te metode, pri prepoznavanju strukture kodirajočih zaporedij, je napovedovanje cepitvenih mest, ki definirajo meje med eksoni in introni (Bendtsen in sod., 2004).

- **Metoda podpornih vektorjev** (SVM*, angl. Support Vector Machine*) – Cilj te metode je najti funkcijo, ki bo uvrstila nabor vhodov v razrede, tako da je prostor med razredi oz. ločitvena meja čim širša. Metoda je zelo robustna in uporabljena na mnogih področjih analize bioloških podatkov, kot npr. analiza transkriptomskih podatkov mikromrež, napoved proteinskih interakcij iz primarnih zaporedij in napoved mest začetka prevajanja (translation initiation site – TIS) (Yang, 2004).

- **Naključni gozdovi** (*angl. random forests*) – Metoda naključnih gozdov temelji na metodi odločitvenih dreves. Pri slednji metodi se na vsakem koraku uporabi pravilo za odločitev med dvema ali več potmi v drevesu. Naključni gozdovi to metodo nadgradijo tako, da s postopkom testa samovzorčenja (*angl. bootstrap*) iz osnovne množice naredijo več novih učnih množic in na podlagi vsake izgradijo odločitveno drevo. Napovedi vseh izgrajenih odločitvenih dreves se na koncu združi v končno napoved. Ena izmed implementacij te metode je na področju napovedovanja glikozilacijskih mest proteinov (Yang in sod., 2010).

- **Genetski algoritem** *(angl. genetic algorithm*) – Je vrsta algoritma, ki preišče velik prostor potencialnih rešitev nekega problema za optimalno rešitev. Zasnovan je na principu Darwinove naravne selekcije, pri čemer začetno populacijo podvrže genetskim konceptom križanja, dedovanja in selekcije ter tako simulira preživetje najuspešnejših potomcev, ki ustvarjajo nove generacije. Metoda ima implementacije na področju

poravnav zaporedij ter napovedi terciarnih struktur proteinov iz primarnih zaporedij (Manning in sod., 2013).

- **Profilna kontekstno neodvisna stohastična gramatika** (SCFG*, angl. Stochastic Context-Free Grammar*) – SCFG metoda je zelo podobna HMM metodi. Od nje se razlikuje v tem, da je linearna pot prehoda stanj zamenjana z drevesom stanj. Originalno izhaja iz področja procesiranja naravnega jezika in deluje na principu generiranja slovnice iz osnovnih znakov (abecede) ter pravil, kako se ti znaki povezujejo. Stohastična različica te metode doda še verjetnost, da se uporabi posamezno pravilo. Najbolj znana implementacija te metode v bioinformatiki je napovedovanje sekundarnih struktur RNA molekul (Knudsen in Hein, 1999; Giegerich, 2011).

Število področij uporabnosti teh modelov stalno narašča, kot tudi programskih implementacij orodij za njihovo uporabo v bioinformatiki. Za izboljševanje napovedi teh modelov pa ni zmeraj pomembna samo izbira ustreznega modela, ampak lahko z združevanjem več modelov dosežemo boljše rezultate kot s posameznim. Tudi ta princip napovedovanja (*angl. ensemble learning*) se postopoma uveljavlja na področju bioinformatike.

2.5.3 Pripis genomskih značilnosti

Proces pripisa genomskih značilnosti je ponavadi sestavljen iz dveh korakov: v prvem koraku se identificira potencialne genomske elemente (odprte bralne okvirje, gene, kodirajoče regije, regulatorne motive) in njihove strukture v genomu, kar imenujemo ***pripis strukturnih značilnosti***. V drugem koraku pa se tem elementom pripiše njihova biološka funkcionalnost (biokemijska funkcija, biološka funkcija, morebitne regulacije in interakcije), kar imenujemo ***pripis funkcionalnih značilnosti*** (Bright in sod., 2009).

Oba koraka pripisa genomskih značilnosti se poskušata izvesti v komplementarnosti dveh pristopov (Petty, 2010):
- Avtomatiziranega pristopa, ki uporablja *in-silico* orodja ter statistične modele za napovedovanje pripisa strukturnih značilnosti;
- Ročnega urejanja (*angl curation*), pri kateri se napovedi iz prvega koraka ročno preverijo ter po potrebi popravijo ali dopolnijo, za kar je ponavadi odgovoren strokovnjak s področja biologije organizma, kateremu se pripisuje značilnosti.

Metoda, ki je precej preprosta in uporabna predvsem pri pripisu strukturnih značilnosti prokariontskih genomov, je **metoda odprtih bralnih okvirjev** (ORF). Ideja te metode je, da so protein-kodirajoči geni sestavljeni iz zaporedij ne-stop kodonov, ki se začnejo s

start kodonom in končajo s stop kodonom. Te okvirje poleg kodirajočih regij sestavljajo tudi signalne regije za transkripcijo in regulatorne regije, kot je npr. promotorska regija. Struktura gena, kjer so odprti bralni okvirji sestavljeni v eno združeno enoto, se pojavlja samo v prokariontskih genomih. Evkariontski geni poleg teh vsebujejo še nekodirajoče intronske regije, ki se ob prepisu izrežejo in niso prisotne v končnih transkriptih. Zaradi tega se lahko proces iskanja genov v prokariontskih genomih poenostavi kar v proces iskanja odprtih bralnih okvirjev. Za pripis značilnosti evkariontskih genomov so primernejše bolj robustne metode od metode iskanja odprtih bralnih okvirjev, kot so npr. v prejšnjem poglavju omenjeni HMM-ji (Cristianini in Hahn, 2006).

V glavnem se metode za pripis značilnosti genomov delijo v 2 glavni kategoriji: *ab-initio,* ki temeljijo na statističnih značilnostih zaporedij in metode na osnovi homologije, ki primerjajo zaporedja z že znanimi in zaporedji, ki imajo pripisane genomske značilnosti (Cristianini in Hahn, 2006). V naslednjih poglavjih bosta predstavljena obe kategoriji.

2.5.3.1 *Ab-initio* napovedi genskih struktur

Programi za *ab-initio* napovedovanje genov, ki podajo napovedi samo na osnovi gole genomske sekvence, so pomemben del procesa pripisa značilnosti genomu. Poglavitni sestavni del teh programov so že prej omenjeni statistični napovedni modeli (npr. skriti markovski modeli – HMM), ki so pripravljeni tako, da poiščejo določene značilnosti genov, kot so npr. eksone, cepitvena mesta, start in stop kodone, introne in nekodirajočo DNA, ki se nahaja v medgenskem prostoru. Posplošen skriti markovski model (GHMM, *angl. General Hidden Markov Model*) je še posebej prilagojen tip HMM, ki lahko s svojim modelom napove genske strukture z dolžinami intronov in eksonov, prilagojenih na vnaprej znane porazdelitve njihovih dolžin za preučevan ali vsaj soroden genom. Vhodni podatki za programe, ki implementirajo te modele, morajo biti samo zaporedja, ki predstavljajo genom. Kot rezultat izvajanja teh programov dobimo predvidene koordinate genskih struktur za preučevani genom (Haas in sod., 2011).

Primer programa, ki deluje po principu *ab-initio* in ne zahteva predhodnega učenja statističnega modela, je GeneMarkS (Besemer in sod., 2001). Ta program uporablja iterativno metodo za iskanje genov v prokariontski DNA, pri kateri se uporabijo hevristični markovski modeli protein-kodirajočih regij in Gibbsovo vzorčenje s poravnavo več zaporedij, ki je implementirano v programu GeneMark.hmm. Slednji proces je del iterativne funkcije, ki ustvari dvo-komponentni statistični model evolucijsko ohranjenih mest v nasproti-točnih (*angl. upstream*) zaporedjih. Ti dve

komponenti sta pozicijsko-frekvenčna matrika (motiv) kodirajočih regij in porazdelitev dolžin medgenskih regij.

Program GeneMarkS iterativno izvaja program GeneMark.hmm in poišče najverjetnejšo (*angl. maximum likelihood*) porazdelitev vhodnega genomskega zaporedja v kodirajoče in ne-kodirajoče regije za določeno iteracijo izvajanja programa. To razdelitev nato uporabi za posodobitev napovedi strukture modelov v naslednji iteraciji. Celoten proces teče do konvergence, ko se razlika v napovedi razdelitev zaporedij v dveh zaporednih iteracijah razlikuje za manj, kakor za neko vnaprej izbrano vrednost. Zaradi tega se lahko GeneMarkS iterativna procedura uporabi na anonimni genomski DNA, brez predhodnega znanja o proteinskih ali rRNA kodirajočih zaporedjih (Besemer in sod., 2001).

2.5.3.2 Metode za pripis genskih struktur na podlagi homologije zaporedij

Avtomatičen postopek pripisa genomskih značilnosti lahko izboljšamo tudi z uporabo statističnih modelov, ki uporabijo predhodno znanje o genomu, na podlagi učenja s homolognimi zaporedji. Ta lahko izvirajo iz homolognih genomskih, transkriptomskih ali proteomskih podatkov, pri čemer je najpogosteje implementirana programska rešitev z uporabo genomskih podatkov. Programi, ki delujejo na tem principu, potrebujejo poleg genomskega zaporedja tudi nabor homolognih genskih zaporedij, ki odražajo podobno strukturo, kot jo želimo napovedati v genskih modelih novega genoma. Veliko programov, ki deluje na tem principu, implementira korak učenja s homolognimi genskimi zaporedji hkrati z napovedovanjem strukturnih modelov. Zelo pomemben korak pred uporabo teh programov pa je priprava preverjenega in ustreznega učnega nabora podatkov znanih struktur kodirajočih regij, ki se uporabi za oceno parametrov za signale cepitvenih mest, kot tudi za porazdelitev dolžin in nukleotidno sestavo eksonov, intronov in medgenskih regij, pomembnih za pravilno napoved struktur genskih modelov (Haas in sod., 2011).

Poleg genomskih podatkov se lahko učni nabor podatkov visoke kvalitete za novi genom pripravi tudi s pomočjo transkriptomskih podatov (RNA-Seq) nekega organizma, iz katerih lahko pridobimo RNA transkripte v polni dolžini ali pa proteomskih podatkov tega organizma, ki podobno kot RNA transkripti, predstavljajo končne produkte izražanja kodirajočih regij. Prednost tega pristopa je, da ne potrebujemo homolognih zaporedij, ampak lahko uporabimo zaporedja, ki izhajajo iz organizma, katerega želimo preučiti (Haas in sod., 2011).
Zaporedja transkriptov, izhajajoča iz istega organizma, katerega genom preučujemo, nam predstavljajo najbolj natančen vir informacij za določitev njegove strukture genov,

ker so zaporedja identična genomskim in natančno določajo meje med introni in eksoni. Ta zaporedja lahko predstavljajo nek del transkripta (EST-ji, sekvenciranje prve in druge generacije) ali pa v najboljšem primeru celoten transkript (FL-cDNA). Slednja zaporedja se težko pridobijo, vendar so za pripis značilnosti genskih struktur najbolj zaželjena, ker v idealnem primeru vsebujejo mesto začetka prepisovanja (transcriptional start site), vse eksone končnega transkripta in poliadenilacijsko cepitveno mesto na 3' koncu. Z natačnimi razcepljenimi poravnavami teh transkriptov na genom pa lahko določimo tudi druge komponente genskih struktur, kot npr. odprt bralni okvir (ORF) in terminalne neprevedene regije (UTR) eksonov (Schadt, 2010).

Za sestavo transkriptov iz RNA-Seq podatkov, pridobljenih z uporabo NGS tehnologije, sta na voljo 2 splošna pristopa: **'mapping first'** strategija, pri kateri se kratki odčitki najprej poravnajo na genom, čemur sledi lokalno sestavljanje poravnav v večje transkriptne strukture; **'assembly first'** strategija, pri kateri se najprej sestavijo zaporedja transkriptov in se ta nato poravnajo na genom, da se lahko določijo strukture genov. Izziv kartiranja milijonov kratkih RNA-seq odčitkov na genom, hkrati z upoštevanjem, da nekateri odčitki prečkajo meje intronov, rešujejo orodja, ki so bila izrecno razvita za uporabo na tem tipu podatkov in uporabljajo strategijo razcepljenih poravnav, kot npr. Tophat, GSNAP in Mapsplice. (Haas in sod., 2011).

Potencialni vir FL-cDNA v prihodnje nam predstavljajo tudi tehnologije sekvenciranja tretje generacije (npr. PacBio SMRT, Oxford NanoPore in druge), katere stremijo k sekvenciranju celotnih molekul, brez vmesnih stopenj drobljenja in sestavljanja kratkih odčitkov (Schadt, 2010).

2.5.3.3 Pripis značilnosti ne-kodirajočih RNA genov

Nekatere RNA molekule ne kodirajo proteinov, ampak namesto tega služijo kot funkcionalni produkti z encimatsko in/ali strukturno vlogo v pomembnih biomolekularnih procesih. Večji razredi ne-kodirajočih RNA genov so vpeti v različne procese, kot npr.: prepisovanje, post-translacijsko mRNA procesiranje in prevajanje. Poznani primeri ne-kodirajočih RNA so tudi npr. ribosomalna RNA (rRNA), ki sestavlja strukturne in funkcionalne komponente ribosomov, prenašalna RNA (tRNA), ki pomaga pri prevodu mRNA v proteine, majhna jedrna RNA (snRNA), ki sodeluje pri cepitvi intronov v pre-mRNA, majhna nukleolarna RNA (snoRNA), ki usmerja biokemijske modifikacije k drugim RNA genom ter mikro RNA (miRNA) in majhna interferenčna RNA (siRNA), ki s svojim delovanjem v splošnem regulirata izražanje določenih tarčnih genov (Haas in sod., 2002).

Računalniški pristopi za pripis značilnosti takšnih nekodirajočih RNA genov se precej razlikujejo od pristopov za iskanje genov, ki kodirajo proteine. Ker so slednji shranjeni z ne-naključnimi kombinacijami kodonov, se lahko ustrezno zaporedje poišče kot zaporedje s statističnimi značilnostmi, ki ustreza znanim genom, za katere vemo, da kodirajo proteine. Informacija v ne-kodirajočih RNA genih nima takšne kodonske strukture, ampak se jo lahko prepozna po obliki sekundarne strukture, ki ustreza zaporedjem in strukturam znanih razredov ne-kodirajočih RNA genov (Gautheret in Lambert, 2001).

Zaporedje in sekundarno strukturo teh ncRNA genov lahko zajamemo s SCFG-ji. Poizvedba s temi je računsko zelo zahtevna in tudi počasna, zato se zavoljo pospešitve analiz uporabi hevristika zadetkov, kjer se najprej zaporedja preišče z algoritmom BLAST, nato pa se počasna SCFG poizvedba osredotoči na genomske regije, ki jih je algoritem BLAST označil kot podobne znanim članom družin ne-kodirajočih RNA (Haas in sod., 2002).

2.5.3.4 Avtomatsko modeliranje genov z uporabo združevalcev dokazov

Ker lahko napovedni modeli predlagajo različne strukture genov na določenem lokusu, je združevanje programov za napovedovanje genov ponavadi najboljši pristop za pripis genskih značilnosti. Metode, ki omogočajo združevanje rezultatov teh programov, se razlikujejo po kompleksnosti: od preproste sheme večinskih predikcij, do kompleksnejših stohastičnih metod, ki uporabljajo ocene za posamično napoved in iz njih točkovanja za doprispevek vsake metode posebej. Primer slednjega je program Maker (Cantarel in sod., 2008), ki združi *ab initio* genske napovedi, RNA-Seq podatke, poravnave proteinov in morebitne ponovitve v genomu ter jih uporabi za napoved genskih struktur. Glavni cilj takšnih metod je doseči nivo natančnosti, ki je enak ali boljši od pripisa značilnosti človeškega skrbnika (ang. *curator*) (Cantarel in sod., 2008).

2.5.3.5 Ročno modeliranje genov z uporabo urejevalca pripisanih genomskih značilnosti

Navkljub uspešno izvedenem avtomatiziranem pripisu genomskih značilnosti je genske modele potrebno preveriti tudi ročno, ker se lahko pojavijo konflikti v podatkih, se uporabi premalo dokazov o strukturi, nastopijo napačne napovedi in pride še do drugih redkejših napak. Z uporabo poravnav homolognih proteinov ali pa poravnav transkriptov lahko strokovni skrbniki (ang. *curators*) predvidijo regije, kjer se nahajajo morebitni konflikti v napovedih in jih odpravijo. Primeri takšnih konfliktov so npr.

nepravilno razcepljeni ali združeni geni, manjkajoči in odvečni eksoni, napačni start in stop kodoni ter nepravilna cepitvena mesta. Te nepravilnosti se pogosto odpravijo v genomskih brskalnikih, ki poleg prikaza omogočajo tudi urejanje pripisanih značilnosti. V urejevalniku takšnih brskalnikov je možno ustvariti modele novih genov, popraviti obstoječe, izbrisati odvečne in - predvsem v primeru evkariontskih organizmov - popraviti meje med introni in eksoni, kjer so prisotne nepravilnosti (Lee in sod., 2013).

Ta pristop je izjemno časovno potraten in zaradi človeškega faktorja tudi podvržen morebitnim napakam, zato zahteva visoko stopnjo predznanja za uspešno izvedbo. Navkljub tem razlogom pa vseeno velja za najboljši pristop za pripis genomskih značilnosti po najvišji stopnji kakovosti (Lee in sod., 2013).

2.5.3.6 Pripis funkcionalnih značilnosti

Ko je končana faza določevanja struktur, je naslednji korak določitev funkcije predvidenim genom. Namen tega koraka je pripis produktov genom na temelju *in silico* karakterizacije s pomočjo homolognih podatkov.

Za določitev funkcije je na voljo več različnih pristopov in podatkovnih baz, ki temeljijo na sekvenčni ali statistični podobnosti podatkov. Primeri funkcionalnih karakterizacij je npr. pripis Gene Ontology (GO) lastnosti, zapisa encimske komisije (EC številke), KEGG metabolne/signalne poti, KOG-homologija, proteinske domene, (PFAM, PRINTS) sekrecijski signali (SignalP) in transmembranske domene (TMHMM). V primeru glivnih genomov pridejo v poštev še kategorije specializiranih funkcij: proteinske kinaze, histidinske kinaze, karbohidrat-aktivirajoči proteini (CAZy), GPI-sidrani proteini, transporterji, sekretorni proteini in efektorji, potencialni patogenski faktorji (PHI-base), gruče genov za sekundarni metabolizem (SMURF), proteaze in transkripcijski faktorji (Haas in sod., 2011).

Za pripis funkcionalnih značilnosti obstajajo tudi alternativni pristopi, ki poskušajo genom pripisati značilnosti na podlagi konteksta, v katerem se pojavljajo, in ne temeljijo zgolj na primerjavi zaporedij ali struktur. Takšna napoved funkcije glede na kontekst je komplementarna napovedim glede na homologijo. Kontekst v tem primeru vsebuje vse tipe povezav med geni in proteini, ki bi lahko nakazovale na funkcionalno interakcijo. Npr. če ima gen A funkcijo X in se gen B funkcionalno povezuje z genom A, potem se preko te povezave lahko sklepa, da ima gen B prav tako vlogo pri funkciji X. Konkretni primeri takšnih povezav so npr. profili proteinskih družin, fuzije domen v multidomenskih proteinih, filogenetski profili (so-pojavljanje genov v genomih), sintenije ter profili genskega izražanja. Geni, katerih produkti so vpleteni v podobnih

funkcijah (npr. tvorijo različne podenote proteina), bodo skupno prisotni ali pa odsotni v določenih genomih in bodo pogosto imeli tudi podobne profile izražanja (kot npr. encimi določene metabolne poti). Na podlagi teh dokazov se lahko pripiše značilnosti tudi genom, ki nimajo na voljo nikakršnih znanih homologov (Huynen in sod., 2000).

2.5.4 Filogenetske analize

Metode filogenetskih analiz nam omogočajo rekonstrukcijo filogenetskih odnosov med taksonomskimi enotami in njihovimi pripadajočimi homolognimi zaporedji. Delijo se v grobem na 2 skupini: tiste, ki rangirajo vsa možna drevesa po nekem kriteriju in tako poiščejo optimalno drevo in tiste, ki izgradijo drevo neposredno iz podatkov (brez eksplicitne funkcije točkovanja). V prvi skupini je najpogostejši pristop iskanje drevesa z najmanjšim številom potrebnih mutacij za razlago podatkov (na podlagi *največje verjetnosti* in drugih metod). Zaradi velikega števila potencialnih dreves lahko te metode porabijo precej časa da poiščejo najboljše drevo, lahko pa da tudi tega ne morejo najti zaradi približkov, ki jih uporabijo za pospešitev iskanja. V vsakem primeru so verjetnostne metode trenutno najbolj priljubljene za globinske filogenetske analize (Cristianini in Hahn, 2006).

Druga skupina vsebuje filogenetske metode, ki so hkrati kriterij in algoritem za izgradnjo dreves in se pogosto začnejo z izračunom parnih oddaljenosti med taksonomskimi enotami. Običajno so zelo hitre in so zaradi tega postale priljubljene v genomskih analizah. Čeprav se statistično ne smatrajo enako dobre kot druge metode, je najbolj znana metoda iz te skupine – algoritem združevanja sosedov *(angl. neighbor-joining - NJ)* precej robustna in natančna (Cristianini in Hahn, 2006).

2.5.5 Analiza evolucijskega pritiska

Najbolj uveljavljena metoda za ta namen je primerjava ne-sinonimnih in sinonimnih zamenjav znotraj gena. Ne-sinonimne mutacije DNA zaporedja so tiste, ki spremenijo zaporedje kodona in njegov aminokislinski produkt, sinonimne pa tiste, ki spremenijo zaporedje kodona, brez da bi spremenile njegovo pripadajočo aminokislino (Yang in Nielsen, 2000).

Iz standardnega genetskega koda je razvidno, da večino sprememb na 3. poziciji kodonov ne spremeni kodirajoče aminokisline. Spremembe na 1. poziciji kodona so lahko včasih sinonimne, medtem ko so spremembe na 2. poziciji vedno ne-sinonimne (Bofkin in Goldman, 2007).

Predpostavka evolucijskih analiz je, da sinonimne zamenjave nimajo nobenih učinkov na funkcijo proteina in posledično tudi na fitnes organizma, medtem ko ne-sinonimne mutacije spremenijo proteinsko zaporedje in s tem lahko tudi učinkujejo na fitnes. Ker sta razmerji ne-sinonimnih in sinonimnih sprememb proporcionalni dejanski hitrosti mutacij, bo edina razlika med njima v številu mutacij, ki obstanejo v populaciji, kar je neposredna posledica naravne selekcije (Suzuki, 2011).

Ker je v vsaki kodirajoči regiji možnih več ne-sinonimnih mutacij kot sinonimnih, se za primerjavo evolucijskega pritiska izračuna število ne-sinonimnih zamenjav na ne-sinonimno mesto (K_a) proti številu sinonimnih zamenjav na sinonimno mesto (K_s) (Suzuki, 2011).

2.5.5.1 Ka/Ks

Koeficient K_a/K_s primerja stopnjo ne-sinonimnih zamenjav na mesto s stopnjo sinonimnih zamenjav na mesto. Ker se sinonimne zamenjave obravnavajo kot funkcionalno nevtralne (tihe), se lahko njihova stopnja zamenjav uporabi kot osnova, proti kateri se lahko interpretira stopnja spreminjanja aminokislin. Relativno veliko število ne-sinonimnih zamenjav lahko nakazuje na delujočo pozitivno selekcijo, ki favorizira nove proteinske strukture (ali pa pojenjanje negativne selekcije proti spremembi proteinov). To nakaže vrednost K_a/K_s večja od 1, medtem ko manjše vrednosti nakazujejo delujočo negativno selekcijo proti spremenljivim mutacijam in posledični ohranitvi proteinske strukture (Vitti in sod., 2013).

Te metode se lahko aplicirajo čez celoten odprt bralni okvir ali pa na določen pod-del zaporedja, ker so lahko različne regije proteina pod različnimi selekcijskimi pritiski. Različni modeli za izračun sinonimnih in ne-sinonimnih stopenj zamenjav upoštevajo različne verjetnosti mutacij (npr. tranzicije so bolj verjetne kot transverzije), kot tudi verjetnosti neopaženih sprememb (npr. več mutacij na istem mestu, ki se izničijo) in razlike v uporabi kodonov (Vitti in sod., 2013).

2.5.5.2 Pristopi za izračun Ka/Ks

Dosedaj se je razvilo že vrsto metod za izračun Ka/Ks koeficienta, ki se v splošnem delijo na 2 razreda: Metode aproksimacije in metode največje verjetnosti (*angl. maximum likelihood*).

Metode aproksimacije sestavljajo 3 koraki:

1) Štetje sinonimnih in ne-sinonimnih mest;
2) Štetje števila sinonimnih in ne-sinonimnih zamenjav;
3) Popravek za večkratne zamenjave.

Metode največje verjetnosti integrirajo evolucionarne lastnosti (izražene v nukleotidnih modelih) v kodonsko-osnovane modele in uporabijo teorijo verjetnosti, da opravijo vse 3 korake hkrati. Te metode uporabljajo različne modele zamenjav in mutacij, osnovanih na različnih predpostavkah o lastnostih zaporedij, kar privede do precej variabilnih rezultatov o evolucijski razdaljah glede na uporabljen model (Zhang in sod., 2006).

Pristopa, ki želita ta proces izboljšati in bolje modelirati lastnosti zaporedja sta: izbor optimalnega modela (*angl. model selection*) in pa povprečenje modelov (*angl. model averaging*). Za ta namen se uporabijo ocenitvene funkcije (npr. AIC in BIC), ki ocenijo kako dobro posamezni modeli opišejo podatke (Zhang in sod., 2006).

Najpreprostejši razred metod aproksimacije ločeno prešteje število sinonimnih in ne-sinonimnih mest v dveh zaporedjih (S_c in A_c) in tudi število sinonimih in ne-sinonimnih razlik med njima (S_d in A_d). S prilagoditvijo števila možnih mest, ki lahko proizvedejo ne-sinonimne in sinonimne spremembe, se lahko izračuna ne-sinonimne in sinonimne zamenjave za vsako mesto, ki so nujne za normalizirano razmerje K_a/K_s (po popravku za večkratne zamenjave) (Cristianini in Hahn, 2006).

Te metode privzamejo, da sta pogostosti transverzij in tranzicij enaki in da ni razlike v uporabi kodonov, medtem ko kompleksnejši algoritmi upoštevajo razlike med pogostostjo tranzicij in transverzij, kot tudi razlike v uporabi kodonov in nukleotidov (Cristianini in Hahn, 2006).

2.5.5.3 Nei in Gojobori algoritem za izračun Ka/Ks

Primer ene izmed osnovnih metod je algoritem Nei-Gojobori, pri katerem se najprej za izračun deleža sinonimnih in ne-sinonimnih zamenjav naredi poravnava homolognih DNA zaporedij (Zhang in sod., 2006). Zaporedji morata biti poravnani na nivoju

kodonov (če so vrzeli v katerem izmed zaporedij, se ti kodoni izključijo iz izračuna), nato sledijo 3 koraki:

Korak 1 - Štetje števila ne-sinonimnih in sinonimnih mest

Vsako protein-kodirajoče zaporedje DNA je zaporedje iz kodonske abecede C, s k-tim kodonom označenim kot c_k. Število sinonimnih mest v k-tem kodonu se označi s $s_c(c_k)$ in ne-sinonimnih mest kot $a_c(c_k) = 1 - s_c(c_k)$. S f_i se označi delež sprememb na i-ti poziciji podanega kodona (i = 1, 2, 3), ki privedejo do sinonimnih sprememb. Količini $s_c(c_k)$ in $a_c(c_k)$ za podani kodon se izračunata kot $s_c(c_k) = Sf_i$ in $a_c(c_k) = (3 - s_c(c_k)) = (3 - Sf_i)$ (Cristianini in Hahn, 2006).

Za DNA zaporedje z *r* kodoni se lahko celotno število sinonimnih in ne-sinonimnih mest poda kot:

$$S_c = \sum_{k=1}^{r} S_c(c_k) \quad \text{... (5)}$$

$$A_c = (3r - S_c) \quad \text{... (6)}$$

Ti dve količini sta lastnosti posameznih zaporedij, natančneje lastnosti specifičnih kodonskih sestav zaporedij (Cristianini in Hahn, 2006). Ker so ti modeli vedno v uporabi pri primerjavi dveh zaporedij, se lahko povprečje S_c in A_c zaporedij, ki se uporabi pri izračunu K_a in K_s koeficientov, definira kot:

$$S_c^{=} = (S_{c1} + S_{c2})/2 \quad \text{... (7)}$$

$$A_c^{=} = (A_{c1} + A_{c2})/2 \quad \text{... (8)}$$

Korak 2 - Štetje ne-sinonimnih in sinonimnih razlik med zaporedji

Količini $s_d(c_k)$ in $a_d(c_k)$ predstavljata število sinonimnih in ne-sinonimnih razlik za k-ti kodonski par v poravnavi. Ko je v poravnavi razlika v samo enem nukleotidu, se lahko preprosto odločimo ali je zamenjava sinonimna ali ne-sinonimna. Ko sta v poravnavi 2 različna nukleotida, potem imamo na voljo 2 možni poti evolucije, ki sta privedli od enega do drugega kodona, odvisno od tega katera mutacija je nastopila prej. Glede na

red zamenjav je lahko določena razlika sinonimna ali pa ne-sinonimna. V takem primeru pripišemo enake verjetnosti obema potema, nato izračunamo količini S_d in A_d za vsako pot ter jih povprečimo. Kadar so v poravnavi različna vsa 3 mesta med kodonoma, potem obstaja 6 različnih poti med njima in v vsaki nastopijo 3 mutacije. Z ozirom na vse te poti in potencialne korake mutacij lahko izračunamo količini S_d in A_d enako kot v primeru razlik v 2 nukleotidih. To pomeni povprečenje S_d in A_d za vsako pot z enakovrednimi utežmi (Cristianini in Hahn, 2006).

Ko imamo izračunane količine S_d in A_d za vsak kodonski par v poravnavi, lahko pridobimo njihove skupne vrednosti za celotno zaporedje z njihovo vsoto (Cristianini in Hahn, 2006). Skupno število sinonimnih in ne-sinonimnih razlik med dvema zaporedjema je:

$$S_d = \sum_{k=1}^{r} S_d(c_k) \quad \text{... (9)}$$

$$A_d = \sum_{k=1}^{r} a_d(c_k) \quad \text{... (10)}$$

pri čemer sta $s_d(c_k)$ in $a_d(c_k)$ količini za s_d in a_d za k-ti kodonski par in r predstavlja število vseh kodonov v primerjavi. Obenem pa je količina $S_d + A_d$ enaka številu vseh nukleotidnih razlik med obravnavanima DNA zaporedjema (Cristianini in Hahn, 2006).

Korak 3 – Popravek za večkratne zamenjave ter izračun Ka in Ks

S količinami, izračunanimi v prejšnjih dveh korakih, se lahko predvidi delež sinonimnih (d_s) in ne-sinonimnih (d_n) razlik preko naslednjih enačb:

$$d_s = S_d / \hat{S_c} \quad \text{... (11)}$$

$$d_a = A_d / \hat{A_c} \quad \text{... (12)}$$

pri čemer $\hat{S}_c$ in $\hat{A}_c$ predstavljata povprečno število sinonimnih in ne-sinonimnih mest v obravnavanih zaporedjih. Za izračun števila sinonimnih (K_s) in ne-sinonimnih (K_a) zamenjav za vsako mesto uporabimo še Jukes in Cantor popravek za d_s in d_a ter tako pridobimo K_a in K_s (Cristianini in Hahn, 2006).

Primer evolucijske analize s pomočjo Ka/Ks koeficientov je bila aplikacija teh metod na analizo efektorskih proteinov patogenih gliv, kateri so bili zaradi hitrega prilagajanja

gostiteljskim odzivom podvrženi povečani stopnji mutacij (de Jonge in sod., 2013). Pri projektu analize genoma *Verticillium tricorpus* so raziskovalci predpostavili, da so bili geni, ki so sodelovali pri procesu patogeneze, podvrženi povečani stopnji pozitivne selekcije in posledično hitrejši evoluciji. Na podlagi te domneve so primerjali razmerje Ka/Ks genov *V. tricorpus* z razmerji Ka/Ks genov drugih *Verticillium* vrst.

Večino genov *V. tricorpus* v tej raziskavi je bilo podvrženih negativni selekciji (Ka/Ks < 1), vendar so imeli geni sekretoma statistično značilno povišana razmerja Ka/Ks v primerjavi z ostalimi geni tega organizma, kar je nakazovalo na pospešeno evolucijo sekretoma (Seidl in sod., 2015).

3 MATERIALI IN METODE

3.1 LABORATORIJSKI DEL

Genomski in RNA-Seq podatki v tem delu izvirajo iz sekvenciranja šestih sevov *V. nonalfalfae s* treh geografskih lokacij:

- Slovenija - T2 in Rec;
- Nemčija - P15 in P55;
- Združeno kraljestvo - 1953 in 1985.

In dveh virulenčnih nivojev, kot je prikazano v preglednici 1**:**

Preglednica 1: Preučevani patotipi
Table 1: Examined pathotypes

Blagi patotip	Letalen patotip
Rec, P55, 1953	T2, P15, 1985

Sekvenciranje je bilo opravljeno na napravah Illumina Genome Analyzer IIx in HiSeq2500 v sekvenatorskem centru IGA Technology Services Srl (Udine, Italija) z velikostjo knjižnic parnih odčitkov velikosti 150 bp. Podatki sekvenciranja so javno dostopni v NCBI SRA arhivu pod bioprojektom PRJNA283258.

Genomski podatki iz analiz so na voljo v naslednjih SRA eksperimentih:

- T2 (SRX1020587, SRX1020589, SRX1020590);
- Rec (SRX1020612, SRX1020613);
- P15 (SRX1021611);
- P55 (SRX1021624);
- 1953 (SRX1021626);
- 1985 (SRX1021650).

RNA-seq podatki za slovenska seva T2 in Rec so na voljo v SRA eksperimentih:

- T2 (SRX1020629);
- Rec (SRX1020679).

DNA šestih glivnih sevov smo izolirali iz kalečih konidijev po postopku izolacije protoplastov (Pantou in Typas, 2005) in CTAB izolacije (Moller in sod., 1992), vključno s tretiranjem z encimom RNAzo. Izolirano DNA smo kvantificirali z NanoVue spektrofotometrom in preverili njeno kakovost z agarozno gelsko elektroforezo.

3.1.1 RT-qPCR dolge ne-kodirajoče RNA (orf414)

Seva T2 in Rec smo nacepili na 1/2 Czapek Dox agarne plošče pri sobni temperaturi. Iz pridobljenega glivnega micelija smo izolirali RNA v treh bioloških ponovitvah (50 mg) za vsak sev z uporabo Spectrum Plant Total RNA Extraction Kit (Sigma-Aldrich) kompleta. Izolacija smo izvedli po protokolu proizvajalca, vključno s tretiranjem z DNAzo na koloni (Sigma-Aldrich). Ovrednotene RNA vzorce smo shranili na -80 °C.

1 µg vsakega RNA vzorca smo reverzno prepisali v cDNA s High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, ZDA), z naključnimi heksamernimi začetnimi oligonukleotidi in shranili na –20 °C. Z Primer Express version 3.0 (Applied Biosystems) na privzetih nastavitvah smo pripravili začetna oligonukleotida za *orf414* ORF414-FOR 5′–ATCCGAGGGAGATGAGACTTCA-3' in ORF414-REV 5′–TCACCTGATCTTTCTTCAACTTCAA-3'. Kot referenco smo uporabil gen za ubikvitin (UBQ), ki smo ga pomnožili z začetnima oligonukleotidoma UBQ-FOR 5′–GACTCGACCTCAAGGGTGAT–3′ ter UBQ-REV 5′–GTCTTCGTGGTGGTATGCAG–3. Ubikvitin je bil izbran za referenčni gen glede na rezultate predhodne raziskave (Duressa in sod., 2013), kjer se je izkazal za najbolj stabilno izraženega pri ekspresijski analizi vrste *V. dahliae*.

Reakcijsko mešanico za qPCR smo pripravili v 10 µl reakcijskem volumnu z 5 µl FAST SYBR Green PCR Master Mix (Applied Biosystems), 2 µl cDNA matrice in 0,6 µM vsakega izmed začetnih oligonukleotidov. FAST amplifikacijo smo izvedli z ABI PRISM 7500 Fast Sequence Detection System napravo (Applied Biosystems, Foster City, ZDA) po naslednjem programu: 95 °C za 20 s, 40 ciklov pri 95 °C za 3 s, čemer je sledilo 60 °C za 30 s in analiza talilne krivulje za potrditev pomnožitve samo enega PCR produkta. Vse vzorce smo pomnožili v tehničnih ponovitvah in jim izračunali vrednost praga (*angl. cycle threshold* (C_t) vrednost), za preučitev nivojev izražanja *orf414*. Primerjava nivojev izražanja je bila opravljena s t-testom za diferencialno izražanje med blagim in letalnim sevom s programom R v3.0.2.

3.1.2 Mitohondrijski dolžinski polimorfizem

Pri poravnavi mitohondrijskih genomov *V. nonalfalfae* in *V. dahliae* smo odkrili *V. dahliae* specifično regijo, katero smo nadaljnje preučili s PCR pomnožitvijo v različnih izolatih vrst *Verticillium* spp. Ohranjeni par začetnih oligonukleotidov mth-for 5'-CCTCACGCTTTTGTAAGTTTACCT-3' in mth-rev 5'-AATTCAAACTCGTTAATACATAGCA-3' smo pripravili s PRIMER3 programsko opremo (Untergasser in sod., 2012).

Začetna oligonukleotida smo uporabili za pomnoževanje zbirke *Verticillium* vrst (96 vzorcev), katero DNA smo imeli na razpolago v laboratoriju katedre (Priloga A):

- *V. albo-atrum* (6 vzorcev);
- *V. alfalfae* (5 vzorcev);
- *V. dahliae* (26 vzorcev);
- *V. isaacii* (3 vzorci);
- *V. nonalfalfae* (49 vzorcev);
- *V. nigrescens* (2 vzorca);
- *V. tricorpus* (2 vzorca);
- *V. longisporum* (2 vzorca);
- *V. nubilum* (1 vzorec).

PCR reakcije smo pripravili v raztopini 20 µl glivne DNA (1:100 razredčena CTAB izolacija), 1x PCR pufra, 2 mM $MgCl_2$, 0,8 mM dNTP raztopine in 0,5U *Taq* DNA polimeraze (Kapa Biosystems). Cikli pomnoževanja so potekali po shemi: 95 °C za 5 min, 5 ciklov pri 95 °C za 30 s, 65 °C za 30 s (pri vsakem ciklu se je temperatura znižala za 1 °C) in 72 °C za 1 min 30 s, čemur je sledilo še 25 ciklov pri istih časovnih in temperaturnih profilih, z izjemo temperature prileganja, ki je bila 55 °C. Produkte pomnoževanja smo ločili na 1,2 % agaroznem gelu.

3.2 BIOINFORMATIČNI DEL

3.2.1 Programska oprema

Pri našem delu smo uporabili več vrst programske opreme za različne namene. V spodnjih Preglednicah 2, 3, 4 in 5 je ta predstavljena po uporabljenih sklopih.

Preglednica 2: Programski jeziki in njihove knjižnice, Linux terminalna orodja
Table 2: Programming languages and their libraries, Linux terminal utilities

Program	Uporabljena verzija
Python	v2.7.6
Python - PIL	v1.1.7
Python - Matplotlib	v1.3.1
Python - Ete2	v2.2.1072
Python - Biopython	v1.63
Bash	v4.3.11
Awk	v4.0.1
Sed	v4.2.2
Perl	v5.18.2
R	v3.0.2
R - DESeq	v1.24.0
GNU Parallel	v20130922

Preglednica 3: Orodja za obdelavo NGS podatkov
Table 3: NGS data analysis tools

Program	Uporabljena verzija
A5	v20140401
Velvet optimiser	v2.2.5
CLC Genomics Workbench	v6.5
Samtools	v0.1.19
Bcftools	v0.1.19
Vcftools	v0.1.11
Bedtools	v2.17.0
Tophat	v2.0.9
Cufflinks	v2.1.1

Preglednica 4: Orodja za poravnave in filogenetske analize
Table 4: Utilities for alignments and phylogenetic analyses

Program	Uporabljena verzija
Muscle	v3.8.31
trimAl	v1.2
gKaKs	v1.3
KaKs Calculator	v1.2
RaxML	v8.1.2
Nucmer	v3.1
orthoMCL	v2.0

Preglednica 5: Orodja za analizo, pripis značilnosti in vizualizacijo zaporedij
Table 5: Utilities for analysing, annotating and visualizing sequences

Program	Uporabljena verzija
SnpEff	v4.0
tRNAscan-SE	v1.23
RNAWeasel	/
Mfannot	/
Blast2GO	v2.8
Transdecoder	v20140704
GenemarkS	v2.5p
Maker2	v2.31.6
Webapollo	v1.11.3
Cusp	v6.6.0.0
ARWEN	v1.2
Circos	v0.66

3.2.2 Strojna oprema

V tem delu smo uporabili spodaj navedeno strojno opremo:

- 8-jedrni Intel Xeon E5-2650 CPU, 32 GB RAM VMware Ubuntu Server 10.04 virtualni strežnik;
- 4-jedrni Intel Core i3-3110M CPU, 6 GB RAM Linux Mint 16 Desktop računalnik.

3.2.3 Razvojno okolje

Ker so raziskave terjale obdelavo večjih količin podatkov, je bilo potrebno v ta namen uporabiti zmogljivo računalniško opremo. Na voljo smo imeli strežnik z 8-jedrnim Intel Xeon E5-2650 procesorjem, 32 GB RAM-a ter zadostno količino pomnilnika za shrambo podatkov, ki je bil nameščen v prostorih Fakultete za računalništvo in informatiko. Na njem smo z uporabo odprtokodnih bioinformatskih programov, programov ukazne vrstice ter Python/Bash/Awk programskih skript sestavili lastne programske cevovode za poganjanje analiz, preko katerih je bilo opravljenih večino dela, z izjemo *de-novo* sestavljanja mitohondrijskega genoma, ki smo ga opravili tudi z uporabo CLC Genomics Workbench programske opreme. Ker je delo potekalo na oddaljenem strežniku, so se zaradi lažje uporabe programski cevovodi zasnovali za poganjanje preko ukazne vrstice.

3.2.4 Opisi analiz

Na strežniku smo namestili potrebno programsko opremo za izvajanje analiz ter tudi knjižnice, ki so bile potrebne za razvoj lastnih programskih skript. Naše opravljene analize se razdelijo v 2 tematska sklopa: analiza mitohondrijskih genomov ter analiza odsekov jedrnega genoma. Ob delu na mitohondrijskih genomih *V. nonalfalfae* smo izvedli *de-novo* sestavljanje, sestavljanje in čiščenje mRNA transkriptov, *ab-initio* pripis značilnosti in pripis značilnosti podprt z integracijo podatkov kodirajočih regij, integracijo pripisanih značilnosti iz več plasti »-omskih podatkov«, ročno urejanje pripisanih značilnosti, filogenetsko analizo, %gc analize ter vizualizacijo končnega mitohondrijskega genoma. Pri vsakem koraku smo preizkusili več bioinformacijskih orodij in večje število kombinacij parametrov. Na podlagi tega smo optimizirali postopke v analizah.

Pri načrtovanju filogenetske analize med sevi *V. nonalfalfae* je prišlo do spremembe v poteku analize, ker so bili *de-novo* sestavljeni mitohondrijski genomi sevov med seboj 100 % identični (potrjeno s ponovnim kartiranjem odčitkov na sestavljene genome) in se je zato prvotno predvidena primerjalna analiza med sevi zamenjala s primerjalno analizo med vrstami.

Analize referenčnega genoma *V. nonalfalfae* so temeljile na kartiranju odčitkov na že sestavljen genom s pripisanimi značilnostmi (Javornik, 2012). Preko njih smo raziskali genomske variante in njihove učinke na različne *V. nonalfalfae* seve. Opravili smo tudi analize gostot eksonov, genov in ponovitev po genomu ter uporabili dva različna pristopa za raziskave evolucijskega pritiska na podlagi Ka/Ks analize.

Analize smo pripravili v obliki programskih cevovodov, pri katerih smo odprto-kodne bioinformatske programe integrirali med seboj z uporabo skript v programskem jeziku Python ter ukazne vrstice v operacijskem sistemu Linux s skriptnima jezikoma Bash in Awk. Programske cevovode ki implementirajo te postopke, smo vstavili v lastno eksperimentalno ogrodje, ki nam je omogočilo reproducibilno ponoviti analize iz surovih podatkov ter strukturirano shranjevanje vhodnih in izhodnih datotek. Hkrati pa nam je omogočilo izrabiti sposobnosti paralelizacije opravkov na enem strežniku s pomočjo orodja GNU-parallel in zaradi modularne sestave obenem vključiti nove tipe analiz po želji. Ta implementacija ogrodja je tudi omogočila shranjevanje sprotnih rezultatov in nadaljevanje analize od zadnjega še uspešno izvedenega koraka; v kolikor se je ta prekinil, poenotene nastavitve globalnih parametrov, sprotno poročanje o poteku analiz, uporabo lokalnih ter datotek na oddaljenih strežnikih in enostavno možnost za razširitev ogrodja ob dodajanju novih analiz. Ker je ogrodje še v razvoju, se je njegova programska koda shranila na spletnem repozitoriju Bitbucket in je na voljo za uporabo ali pa za dodaten razvoj zmogljivosti po dogovoru s skrbnikom repozitorija (Vid Jelen) .

3.2.4.1 Priprava mitohondrijskega genoma

Referenčni mitohondrijski genom *V. nonalfalfae* smo pripravili z *de-novo* sestavljanjem genomskih odčitkov. Soseske, ki smo jih dobili pri tem postopku, smo s programsko opremo Blast+ poravnali na mitohondrijski genom *V. dahliae* (NC_008248.1). Za vseh 6 sevov smo najprej preizkusili različne *de-novo* programe za sestavljanje zaporedij (*angl. assembler*) (A5 (Tritt in sod., 2012), Velvet (Zerbino, 2010), CLC Genomics Workbench) in po procesu optimiziranja parametrov ter večih iteracij izvajanj programov ugotovili, da so mitohondrijski genomi teh sevov medsebojno identični (kar smo potrdili tudi s ponovnim kartiranjem odčitkov na sestavljene mitohondrijske genome). Na podlagi tega smo izbrali enega izmed pridobljenih sestavkov za referenčnega in ga uporabili za vse nadaljnje delo.

Za kvalitetne pripise značilnosti mitohondrijskega genoma smo želeli vključiti različne nivoje "-omskih" podatkov. Zaradi tega smo poleg genomskih podatkov dodali še transkriptomske (RNA-Seq) ter proteomske podatke (proteome s pripisanimi značilnostmi sorodnih *Verticillium* vrst z Genbank repozitorija).

Z uporabo RNA-Seq podatkov smo s programskim cevovodom, v katerem smo uporabili programa Tophat (Kim in sod., 2013) in Cufflinks (Roberts in sod., 2011) najprej *de-novo* sestavili transkriptom ter nato v pridobljenih transkriptih poiskali kodirajoče regije z orodjem Transdecoder (Haas, 2015). Pridobljene podatke smo nato

uporabili v programu Maker2 (Holt in Yandell, 2011), ki združuje *ab-initio* napovedovalca kodirajočih regij GeneMark (Borodovsky in McIninch, 1993), poravnave 871 V*erticillium* spp. glivnih mitohondrijskih proteinov (NCBI Entrez query: "fungi[Organism] AND verticillium AND mitochondrial", September 2014) iz 'nr' Genbank zbirke z orodjem Exonerate (Slater in Birney, 2005) ter sestavljenimi transkripti na podlagi RNA-Seq podatkov. Podatke RNA-Seq kartiranja s programom Bowtie2 (Langmead in sod., 2009) smo uporabili za analizo diferencialnega izražanja genov na mitohondrijskem genomu. Ta analiza je potekala v okolju R s pomočjo programske knjižnice DESeq (Anders in Huber, 2010) in lastnih Python skript.

Za določevanjem RNA struktur v mitohondrijskem genomu smo združili rezultate orodij RNA weasel, Mfannot in tRNAscan-SE (Lowe in Eddy, 1997), s katerimi smo napovedali različne razrede prisotnih RNA struktur in intronov. Določili smo tudi uporabo kodonov s programom cusp iz paketa EMBOSS (Rice in sod., 2000) in analizirali %gc vsebnost ter %gc odstopanje (*angl. skew*) z lastnimi Python programskimi skriptami. Avtomatično določene genske modele in RNA strukture smo pred pripisom funkcionalnih značilnosti še ročno preverili in popravili na lastnem Webapollo (Lee in sod., 2013) strežniku in v njem pripravili tudi končne genske modele. Pripis funkcionalnih značlinosti smo izvedli z uporabo programa Blast2GO (Conesa in sod., 2005), dodatno pa smo še sekundarne strukture tRNA molekul določili s programom ARWEN (Laslett in Canbäck, 2008) in jih vizualizirali s programom VARNA (Darty in sod., 2009). Programe smo izvajali s privzetimi nastavitvami, z izjemo nastavitve genetskega koda na mitohondrijski kod. Vizualizacijo končnega mitohondrijskega genoma smo pripravili z orodjem Circos (Krzywinski in sod., 2009).

Za preiskavo morebitnih homologov in potencialnih funkcij dolge ne-kodirajoče RNA *o*rf414 smo njeno karakterizacijo izvedli s spletnim strežnikom BLAST v okviru NCBI z uporabo BLASTn poizvedb po 9 nukleotidnih podatkovnih bazah (nucleotide collection (nr/nt), reference RNA sequences (refseq_rna), reference genomic sequences (refseq_genomic), refseq representative genomes (refseq_representative_genomes), NCBI Genomes (chromsome), expressed sequence tags (est), genomic survey sequences (gss), high throughput genomic sequences (HTGS), transcriptome Shotgun Assembly (TSA) sequences (tsa_all)) in BLASTX poizvedbo po ne-presežni zbirki proteinskih zaporedij (nr). Grafični prikaz mitohondrijskega genoma s pripisanimi značilnostmi smo pripravili z orodjem Circos. Končno zaporedje mitohondrijskega genoma *V. nonalfalfae* smo shranili v zbirki Genbank pod pristopno številko KR704425.

3.2.4.1.1 Filogenetska analiza mitohondrijskih genomov

Izbrali smo nabor javno dostopnih glivnih mitohondrijskih genomov s popolno pripisanimi značilnostmi iz širše taksonomske skupine glivnih vrst, vključno s 15 patogenimi glivami iz taksonomske skupine Pezizomycotina ter 3 kvasovkami, ki so predstavljale skupino askomicet in enim bazidiomicetnim patogenom, ki je bil namenjen kot izhodna skupina (*angl. outgroup*) pri filogenetski analizi (Preglednica 1).

Preglednica 6: Izbrani organizmi za filogenetsko analizo
Table 6: Organisms selected for the phylogentic analysis

Genbank ID	Taksonomija	Ime	Velikost genoma (bp)	Povprečna GC vsebnost	# tRNA
NC_023540.1	*Glomerellales*	*Colletotrichum lindemuthianum*	36.957	30,88 %	28
NC_001329.3	*Sordariales*	*Podospora anserina*	100.314	30,06 %	27
NC_023127.1	*Helotiales*	*Rhynchosporium orthosporum*	49.539	28,80 %	29
NC_008068.1	*Hypocreales*	*Metarhizium anisopliae*	24.673	28,40 %	24
KC683708.1	*Sordariales*	*Neurospora crassa*	64.840	36,13 %	28
NC_010222.1	*Capnodiales*	*Zymoseptoria tritici*	43.964	31,94 %	27
NC_004514.1	*Hypocreales*	*Lecanicillium muscarium*	24.499	27,15 %	25
NC_016680.1	*Hypocreales*	*Fusarium solani*	62.978	28,88 %	25
NC_025200.1	*Helotiales*	*Sclerotinia borealis*	203.051	32,01 %	31
NC_017930.1	*Hypocreales*	*Fusarium oxysporum*	34.477	30,98 %	25
NC_007445.1	*Eurotiales*	*Aspergillus niger*	31.103	26,90 %	25
NC_001326.1	*Schizosaccharomycetales*	*Schizosaccharomyces pombe*	19.431	30,09 %	25
NC_009493.1	*Hypocreales*	*Fusarium graminearum*	95.676	31,84 %	28
NC_008248.1	*Glomerellales*	*Verticillium dahliae*	27.184	27,32 %	25
NC_023268.1	*Hypocreales*	*Acremonium chrysogenum*	27.266	26,54 %	26
NC_001224.1	*Saccharomycetales*	*Saccharomyces cerevisiae*	85.779	17,11 %	24
NC_018046.1	*Saccharomycetales*	*Candida albicans*	33.928	31,74 %	24
NC_022835.1	*Hypocreales*	*Metacordyceps chlamydosporia*	25.615	28,28 %	22
NC_008368.1	*Ustilaginales*	*Ustilago maydis*	56.814	31,20 %	23
KR704425	*Glomerellales*	*Verticillium nonalfalfae*	26.139	26,92 %	26

Aminokislinska zaporedja 14 ohranjenih mitohondrijskih proteinov iz teh vrst, vključno s štirimi citokrom c oksidaznimi podenotami (cox1, cox2, cox3 and cob), tremi ATP sintaza podenotami (atp6, atp8 and atp9) in sedmimi NADH dehidrogenaza podenotami (nad1, nad2, nad3, nad4, nad4L, nad5 and nad6) smo poravnali s programom Muscle (Edgar, 2004) in jim odstranili slabo poravnana aminokislinska zaporedja s programom trimAl (Capella-Gutiérrez in sod., 2009). Poravnave smo nato konkatenirali in uporabili v izgradnji filogenetskega drevesa s programom RAxML (Stamatakis, 2014). Pri tem smo uporabili avtomatizirano izbiro proteinskega modela glede na verjetnost fiksnega, smiselnega drevesa in GAMMA model stopnje heterogenosti. Končno filogenetsko drevo z največjo verjetnostjo smo naredil z uporabo RAxML rapid bootstrap algoritma s 100 približki (*angl. approximation*) in končnim iskanjem največje verjetnosti. Dobljeno filogenetsko drevo in vrstni red proteinov smo vizualizirali z uporabo programskega jezika Python in knjižnice ETE2.

3.2.4.1.2 Določevanje dolžinskega polimorfizma

Sestavljen *V. nonalfalfae* mitohondrijski genom smo poravnali na *V. dahliae* mitohondrijski genom (Genbank NC_008248.1) z uporabo programa Nucmer (Kurtz in sod., 2004). Dodatno smo kartirali tudi sekvenčne podatke 6 *V. nonalfalfae* sevov na referenčni *V. dahliae* mitohondrijski genom z uporabo 'Map reads to reference' orodja v programski opremi CLC Genomics Workbench.

3.2.4.2 Analiza odsekov jedrnega genoma

Ob analizi jedrnega genoma smo se osredotočili na preučevanje variant, gostotnih porazdelitev eksonskih in ponovitvenih regij ter na analizo evolucijskega pritiska na kodirajoče regije.

3.2.4.2.1 Določevanje variant

Za izhodišče nadaljnjih raziskav smo najprej preučili variante (insercije, delecije, zamenjave) v referenčnem genomu T2. Te smo določili s postopkom kartiranja parnih genomskih odčitkov na posamezne genome z orodjem CLC Genomics Workbench. Za nadaljnje analize smo izvozili BAM datoteke kartiranj.

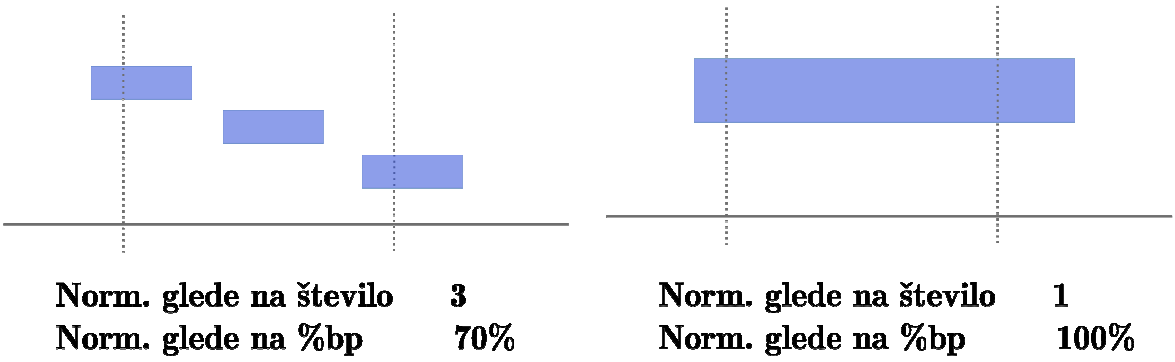
Te datoteke so služile kot temelj za naš programski cevovod, ki je implementiral analizo variant z naborom orodij Samtools, Bcftools (Li in sod., 2009) in Vcfutils (Danecek in

sod., 2011). S temi orodji smo določili in statistično ovrednotili variante ter njihove genotipe in jih nato še dodatno filtrirali glede na najmanjšo (10) in najvišjo (1.000.000) pokritost položaja v genomu ter z minimalno RMS (root-mean-square) kvaliteto kartiranja za SNP-e (20). Iz pridobljenih variant smo nato odfiltrirali lažno pozitivne heterozigotne variante, ker je naš preučevani organizem homozigoten. Končnemu naboru filtriranih variant smo nato pripisali značilnosti glede na njihov potencialen učinek na kodirajoče elemente z orodjem snpEff (Cingolani, 2012) in jih uporabili kot osnovo za nadaljnjo filogenetsko analizo.

3.2.4.2.2 Preučitev gostot eksonskih regij, gostote variant in ponovitev v genomu

Potencialno zanimive regije (anomalije, vroče točke) in gostotne porazdelitve določenih genomskih značilnosti (eksonov, ponovitev, snp-jev) smo analizirali z drsnim oknom velikosti 10 kbp. Normalizacijo vrednosti znotraj okna pri eksonih in ponovitvah smo izvedli kot odstotek baznih parov okna, ki pripada posameznim genomskim značilnostim. Ta postopek se razlikuje od pogosto uporabljenega, kjer se vrednosti normalizirajo glede na število genomskih značilnosti znotraj drsnega okna, kar smo uporabili pri normalizaciji gostote snp-ov. Pri tem se daje večjo težo mestom v genomu, kjer se pojavlja veliko število majhnih genomskih značilnosti in lahko tudi izpusti mesta, kjer je teh malo.

Spodnja slika 2 prikazuje različne rezultate, ki jih pridobimo z obema pristopoma. Pri normalizaciji glede na število genomskih značilnosti dobimo vrednosti, ki nakazujejo prisotnost več značilnosti znotraj drsnega okna, ne nosijo pa informacij o njihovi velikosti. Pri normalizaciji glede na % bp dobimo vrednosti, ki nakazujejo na skupno velikost vseh genomskih značilnosti znotraj drsnega okna. Oba pristopa nosita informacije o gostotni porazdelitvi, ampak glede na cilj analize se izbere pristop normalizacije, ki je zanjo najbolj ustrezen.



Slika 2: Razlika v pristopih normalizacij znotraj drsnega okna
Slika predstavlja genomske značilnosti (eksone, ponovitve ali variante na y-osi) vzdolž genoma (x-os). Na levi strani je prikazana normalizacija večih genomskih značilnosti znotraj drsnega okna, na desni strani pa je prikazana samo ena, ki pokriva celoteno drsno okno. Pri normalizacijskem postopku glede na število bo končna vrednost višja kot glede na % pokritosti okna pri primeru na levi strani in obratno na desni strani. Za ustrezen pristop normalizacije se odločimo glede na cilj analize.

Figure 2: Difference in normalization approaches inside a sliding window
The figure represents genomic features (exons, repeats or variants on the y-axis) along the genome (x-axis). Normalization of several genomic features within a sliding window is shown on the left side and a single feature spanning the entire sliding window is shown on the right side. The final values will be higher by normalizing according to the number of features like seen on the left side, and vice versa on the right side. The appropriate normalization procedure is chosen according to the goal of the analysis.

Za izvedbo analize porazdelitev variant smo uporabili pristop z drsnim oknom, katerega smo implementirali z uporabo lastnih programskih skript ter uporabo programov Vcftools in Bedtools. Preučili smo tudi porazdelitve gostot eksonskih regij ter regij s ponovitvami po podobnem postopku z drsnim oknom. Končne rezultate analiz porazdelitev gostot smo vizualizirali z uporabo knjižnice Matplotlib in lastne Python skripte.

3.2.4.2.3 Analiza evolucijskega pritiska na kodirajoče regije

Za analizo evolucijskega pritiska preko izračuna Ka/Ks koeficientov smo uporabili 2 različna pristopa:

1) Izračun koeficientov med vrstami z uporabo gKaKs skripte (Zhang in sod., 2013), ki omogoča izračun Ka/Ks koeficientov vseh kodirajočih zaporedij v genomu. Pri tem postopku smo poravnali prevedena genska zaporedja referenčnega seva na genomska zaporedja drugih *Verticillium* vrst.

2) Izračun koeficientov med vrstami preko ortolognih genov, ki smo jih pridobili z uporabo programa orthoMCL (Li in sod., 2003). Kot vhodne podatke smo uporabili proteoma *V. alfalfe* (VaMS102) in *V. dahliae* (VdLs17).

Končni izračun Ka/Ks koeficientov je pri obeh metodah temeljil na Nei-Gojobori (NG) in Yang-Nielsen (YN) metodah, ki sta implementirani v programu KaKs Calculator (Zhang in sod., 2006). Razlika med tema metodama je, da YN pri izračunu upošteva razliko med stopnjama tranzicij/transverzij in neenakomerne frekvence porazdelitev kodonov.

3.2.4.2.4 Filogenetska analiza na podlagi variant v jedrnih genomih

Pri tem postopku smo za vsak sev posebej konkatenirali homozigotne SNP variante, ki so po filtriranju imele dovolj dobro kakovost kartiranja ter so bile statistično značilno podprte in jih poravnali s posebej pripravljeno Python skripto. Vsak položaj v poravnavi je lahko izviral iz nekega seva ali pa bil dopolnjen s praznimi mesti, v kolikor določen sev ni imel SNP variante na tistem položaju.

Filogenetsko analizo jedrnih genomov na podlagi variant smo, podobno kot pri filogenetski analizi mitohondrijev, opravili z uporabo orodja RAxML. Pri tem postopku smo uporabili GTR model nukleotidnih zamenjav in privzeli poenoteno stopnjo heterogenosti med mesti. Končno filogenetsko drevo z največjo verjetnostjo smo naredil z uporabo RAxML rapid bootstrap algoritma s 100 približki in končnim iskanjem največje verjetnosti.

3.2.4.2.5 Obogatitvena analiza GO pojmov

Genom, pridobljenim s Ka/Ks analizo, smo preverili obogatitev GO pojmov na podlagi Fisherjevega eksaktnega test-a s popravki za testiranje več domnev hkrati z orodjem goatools (Haibao in sod., 2015).

4 REZULTATI

4.1 MITOHONDRIJSKI GENOM V. nonalfalfae

Končen mitohondrijski genom V. nonalfalfae je krožna DNA molekula z velikostjo 26.139 bp in s povprečno GC vsebnostjo 26,92 %. Trije uporabljeni programi za sestavljanje zaporedij so proizvedli enako mitohondrijsko zaporedje za vseh 6 NGS naborov odčitkov. Tudi ponovno kartiranje NGS odčitkov nazaj na sestavljena mitohondrijska zaporedja in sledeča analiza variant nista pokazala polimorfnih regij, navkljub visoki pokritosti (v povprečju med 499 in 10.688x) za vseh 6 naborov podatkov (Preglednica 7). Po velikosti je *V. nonalfalfae* mitohondrijski genom primerljiv s sorodnim *V. dahliae* mitohondrijskim genomom, ki je objavljen na Genbank repozitoriju (NC_008248.1) in ima velikost 27.184 bp ter povprečno GC vsebnost 27,32 %. Nucmer poravnava obeh zaporedij je pokazala 98,15 % identičnost zaporedij z 99,35 % poravnanimi bazami *V. nonalfalfae* mitohondrija na *V. dahliae*.

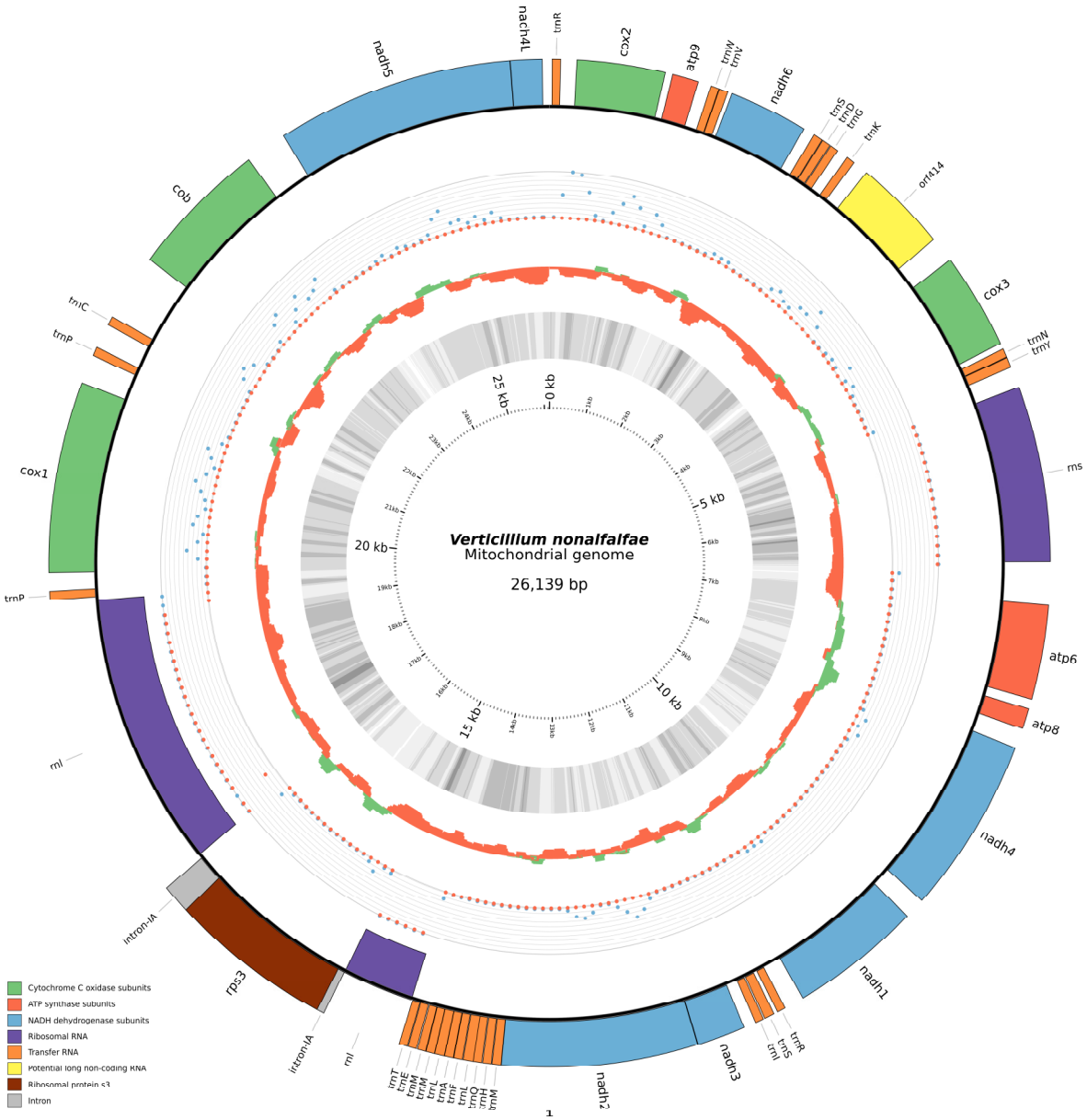
Preglednica 7: Rezultati kartiranja mitohondrijskih odčitkov na sestavljene mitohondrijske genome za 6 uporabljenih NGS naborov podatkov
Table 7: Results of mitochondrial reads mapping to the assembled mitochondrial genomes for all 6 used NGS data sets

Statistka/Sev	Rec	T2	1953	1985	P55	P15
Št. vseh odčitkov	97.295	2.040.416	203.690	761.897	250.232	351.890
Št. parnih odčitkov	84.816	1.825.642	191.878	703.388	236.896	321.930
Celotna dolžina odčitkov (bp)	13.039.773	279.682.422	27.823.727	106.099.232	34.619.109	48.297.310
Najmanjša pokritost (X)	7	65	9	42	17	12
Največja pokritost (X)	1.028	18.987	2.467	8.243	2.853	4.016
Povprečna pokritost (X)	499	10.699	1.064	4.059	1.324	1.847

4.1.1 Pripisane značilnosti mitohondrija

Mitohondrijski genom *V. nonalfalfae* vsebuje celoten nabor 14 ohranjenih mitohondrijskih protein-kodirajočih genov. Ti predstavljajo podenote elektronske prenašalne verige kompleksa I (*nad1, nad2, nad3, nad4, nad4L, nad5, nad6*), kompleksa III (*cox1, cox2, cox3*), kompleksa IV (*cob*) in podenote ATP sintaze (*atp6, atp8, atp9*). Poleg njih se na mitohondrijskem genomu nahajajo tudi: majhna rRNA podenota, velika rRNA podenota, ki vsebuje ribosomalni protein S3 znotraj introna tipa

IA, 26 tRNA-kodirajočih genov in potencialna dolga ne-kodirajoča RNA, ki smo jo poimenovali *orf414* (Slika 3).



Slika 3: Genetska mapa *Verticillium nonalfalfae* mtDNA
Koncentrični krogi od zunaj navznoter predstavljajo različne sloje. Prvi zunanji sloj predstavlja kodirajoče gene *V. nonalfalfe* mitohondrijskega genoma. Smer označb (navzven, navznoter) predstavlja verigo, na kateri je gen prisoten. Naslednji sloj predstavlja število odčitkov RNA-Seq kartiranja iz 2 različnih patotipov *V. nonalfalfae* (blag-modra, letalen-rdeča). Sloja sta pripravljena na podlagi Bowtie2 kartiranih RNA-Seq odčitkov in prikazana kot število odčitkov v 100 bp odsekih. Za normalizacijo teh podatkov se je uporabila DESeq metoda. Ker število rRNA odčitkov odločno presega preostale profile izražanja, se je število odčitkov na grafu omejilo pri 17,5 % najštevilčnejših profilov (ne-omejen graf bi prikazal izražanje samo pri rRNA), za vizualizacijo regij z nizkim izražanjem. Naslednja sloja predstavljata GC-odstopanje in GC-vsebnost. GC-odstopanje (*angl. skew*) [(G-C)/(G+C)] ponazarja relativno število citozinov v primerjavi z gvaninom in se pogosto uporablja za opis verižno-specifične pristranskosti nukleotidne sestave. Ta sloj je predstavljen kot histogram z izračunanimi gc-odstopanja koeficienti v 250 bp drsnih

Slika 3 se nadaljuje

oknih. Zelene regije predstavljajo okna, kjer je koeficient večji od 0, rdeče regije pa predstavljajo okna, kjer je koeficient manjši od 0. Sosednja sivinska toplotna mapa GC-vsebnosti predstavlja 100 bp drsna okna z izračunanimi gc-vsebnostmi. Regije v GC sivinski toplotni mapi so obarvane v temnejšem odtenku, ki predstavlja višjo gc-vsebnost in svetlejšem odtenku, kateri predstavlja nižjo gc-vsebnost. Oba sloja izražata podoben vzorec kot se pojavlja pri drugih sordariomicetah in sta se uporabila za iskanje anomalij v GC vsebnosti, ki bi lahko nakazovale na vdor heterologne DNA. Takšnih anomalij v naših podatkih nismo zaznali. Analizo kumulativnega GC-odstopanja smo uporabili tudi za iskanje lokusov začetka replikacije in terminiranja replikacije, vendar ju s to analizo nismo uspeli določiti.

Figure 3: Genetic map of the *Verticillium nonalfalfae* mtDNA
The concentric circles from the outside inwards represent different tracks. The outermost track represents the coding features of the *V. nonalfalfae* mitochondrial genome. The direction of the highlights (inward, outward) represents the strand in which the feature is present. The next track represents read counts of RNA-Seq mapping from 2 different pathotypes (mild-blue, lethal-red) of *V. nonalfalfa*e. The tracks are made of Bowtie2 [56] mapped RNA-Seq reads and shown as counts per 100 bp bins. The counts were normalized with the DESeq method [57]. Because of rRNA counts overwhelming the remaining expression profiles, the count number on the graph was capped below 17,5 % of the top count profiles (a non-capped graph would show only rRNA expressed), in order to enable visualization of low-expressed regions. Following this track are GC-skew and GC-content tracks, respectively. GC-skew [(G-C)/(G+C)] reflects the relative number of cytosine to guanine and is often used to describe the strand-specific bias of a nucleotide composition. The GC-skew track is shown as a histogram of 250 bp sliding windows with calculated gc-skew coefficients. Green regions represent windows for which the coefficient is larger than 0 and red regions windows for which the coefficient is smaller than 0. The neighbouring grayscale heatmap of the GC-content track represents 100 bp sliding windows with calculated gc-contents. Regions in the GC content heatmap are shaded in gray, where darker gray represents higher gc-content and lighter gray represents lower gc-content. The two tracks show a similar pattern to other *Sordariomycetes* and were also used to scan for anomalies in GC content, which could indicate the introduction of heterologous DNA. No anomalies indicating such an event were detected. The cumulative GC skew analysis was also used to try and find the origins of replication and termination of replication loci (data not shown) but we could not determine them with this analysis.

Vsi mitohondrijski geni so zapisani na isti verigi, z izjemo dolge rRNA podenote, ki je zapisana na nasprotni verigi (Slika 3, Preglednici 8 ter 9).

Preglednica 8: Kodirajoči geni mitohondrijskega genoma *V. nonalfalfae*
Table 8: *V. nonalfalfae mitochondrial genome coding genes*

Gen	Začetna pozicija	Končna pozicija	Dolžina (bp)	Veriga	Začetni kodon	Končni kodon
cox1	19.528	21.129	1.602	-	AUG	UAA
cox2	228	971	744	-	AUG	UAA
cox3	3.855	4.664	810	-	AUA	UAA
cob	22.290	23.459	1.170	-	AUG	UAA
nad1	9.760	10.863	1.104	-	AUG	UAA
nad2	11.820	13.472	1.653	-	AUG	UAA
nad3	11.409	11.816	408	-	AUG	UAA
nad4	8.100	9.584	1.485	-	AUG	UAA
nad4L	25.812	26.078	267	-	AUG	UAA
nad5	23.809	25.809	2.001	-	AUG	UAA
nad6	1.552	2.217	666	-	AUG	UAG
atp6	6.884	7.675	792	-	AUG	UAA
atp8	7.760	7.930	171	-	AUG	UAA
atp9	1.034	1.255	222	-	AUG	UAA
rps3	15.079	16.452	1.374	-	AUA	UAA

Preglednica 9: rRNA in *orf414* mitohondrijskega genoma *V. nonalfalfae*
Table 9: rRNA and *orf414* of the *V. nonalfalfae* mitochondrial genome

RNA	Začetna pozicija	Končna pozicija	Dolžina (bp)	Veriga
rns	5.049	6.525	1.477	-
rnl	14.354	15.014	660	+
rnl	16.734	19.266	2.533	+
orf414	2.850	3.638	789	-

Določili smo tudi domnevne sekundarne strukture tRNA molekul. Vse izmed njih se lahko zvijejo v običajno deteljno strukturo, ki jo sestavlja akceporsko steblo, D-zanka, antikodonska zanka, antikodon in TψC zanka. Pet od vseh tRNA molekul ima še dodatno variabilno zanko: tRNA-Ser(GCU), tRNA-Tyr(GUA), tRNA-Ser(UGA), tRNA-Leu(UAG), tRNA-Ser(UAA). Te molekule so grafično predstavljene na sliki v prilogi (Priloga B).

4.1.2 Uporaba kodonov

Mitohondrijskemu genomu smo določili statistiko njegove uporabe kodonov (Slika 3, Priloga C, Priloga D) in analizirali njegove gene z ozirom na začetne in končne kodone (Preglednica 8). Med predvidenimi geni je najbolj pogost začetni kodon 'AUG', pri čemer le gena *cox3* in *rps3* uporabljata alternativni 'AUA' začetni kodon. Najbolj pogost končni kodon je 'UAA', kateremu sledita dva alternativna končna kodona s podobno frekvenco pojavljanja. 26 tRNA genov (Preglednica 10) kodira 20 od 22 običajnih aminokislin (manjkata Ornitin in Selenocistein), z anti-kodoni ACG;UCU (Arginin), GCU;UGA (Serin), CAU;CAU (Metionin) in UAG;UAA (Levcin), ki so prisotni v več kot 1 kopiji.

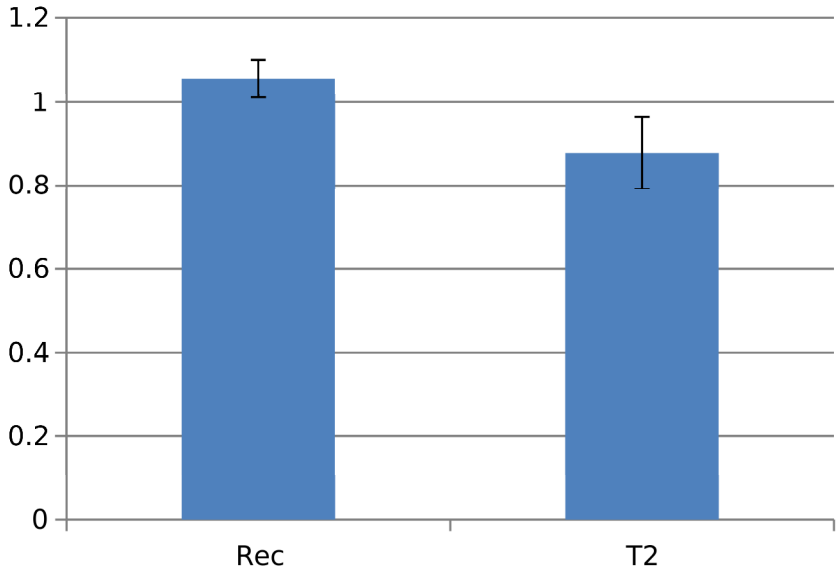
Preglednica 10: tRNA kodirajoči geni mitohondrijskega genoma *V. nonalfalfae*
Table 10: tRNA coding genes of the *V. nonalfalfae* mitochondrial genome

tRNA	Anti kodon	Začetna pozicija	Končna pozicija	Veriga
Arginin	ACG	22	92	-
Triptofan	UCA	1.366	1.437	-
Valin	UAC	1.444	1.515	-
Serin	GCU	2.315	2.394	-
Asparagin	GUC	2.400	2.472	-
Glicin	UCC	2.480	2.550	-
Lizin	UUU	2.647	2.719	-
Asparagin	GUU	4.713	4.783	-
Tirozin	GUA	4.788	4.872	-
Arginin	UCU	11.033	11.103	-
Serin	UGA	11.137	11.221	-
Izolevcin	GAU	11.234	11.305	-
Metionin	CAU	13.477	13.549	-
Histidin	GUG	13.556	13.629	-
Glutamin	UUG	13.636	13.708	-
Levcin	UAG	13.710	13.792	-
Fenilalanin	GAA	13.799	13.871	-
Alanin	UGC	13.878	13.949	-
Levcin	UAA	13.951	14.033	-
Metionin	CAU	14.035	14.107	-
Metionin	CAU	14.111	14.181	-
Glutamat	UUC	14.191	14.263	-
Treonin	UGU	14.270	14.340	-
Prolin	UGG	19.299	19.370	-
Prolin	UGG	21.370	21.441	-
Cistein	GCA	21.648	21.724	-

4.1.3 Karakterizacija in profil izražanja dolge ne-kodirajoče RNA - *orf414*

Karakterizacijo zaporedja *orf414* (2.850..3.638, dolžina 789 bp) smo izvedli z uporabo algoritma BLASTn in BLASTx za iskanje po NCBI podatkovnih bazah. Nukleotidne primerjave so pokazale, da je ta regija specifična za rod *Verticillium,* ker so se statistično značilni zadetki ($e>10^{-5}$, >50 % preiskovanega zaporedja) nahajali le v zaporedjih *V. dahliae*. Večina od njih je imela poravnave v mitohondriju *V. dahliae*, z 96 % identičnostjo in 100 % pokritostjo preiskovanega zaporedja, z izjemo chr2 genomskega zaporedja v nt podatkovni bazi za JR2 *V. dahliae* sestavek, ki ima 93 % identičnost in 88 % pokritost preiskovanega zaporedja. Samo ena statistično značilna podobnost drugi vrsti se je našla v EST odseku, pri cDNA klonu *Phytophtora infestans,* ki je imela 95 % identičnost preko 56 % zaporedja *orf414* (Priloga E). Primerjava s proteinsko podatkovno bazo (BLASTx) ni pokazala nobenih statistično značilnih zadetkov.

Preverili smo tudi *in-vivo* profil izražanja *orf414* v dveh slovenskih glivnih sevih iz različnih patotipov (Rec in T2) z RT-qPCR analizo, pri kateri smo uporabili tri biološke ponovitve. Izražanje *orf414* v blagem sevu je za faktor 1.2 večje od izražanja v letalnem sevu (Slika 4). Vrednosti izražanja smo normalizirali z nivojem izražanja ubikvitina (*ubq*) kot reference in smo jih primerjali med sevi z neparnim t-testom. Povprečna razlika v nivojih izražanja *orf414* med blagim sevom (povprečje = 1,056, standardni odklon = 0,045, standardna napaka = 0,032) in letalnim sevom (povprečje = 0,878, standardni odklon = 0,086, standardna napaka = 0,043) je bila določena s p-vrednostjo 0,05766 in s 95 % intervalom zaupanja [-0,009, 0,365], na podlagi česar lahko sklepamo, da je izražanje *orf414* med sevi na podobnih nivojih.

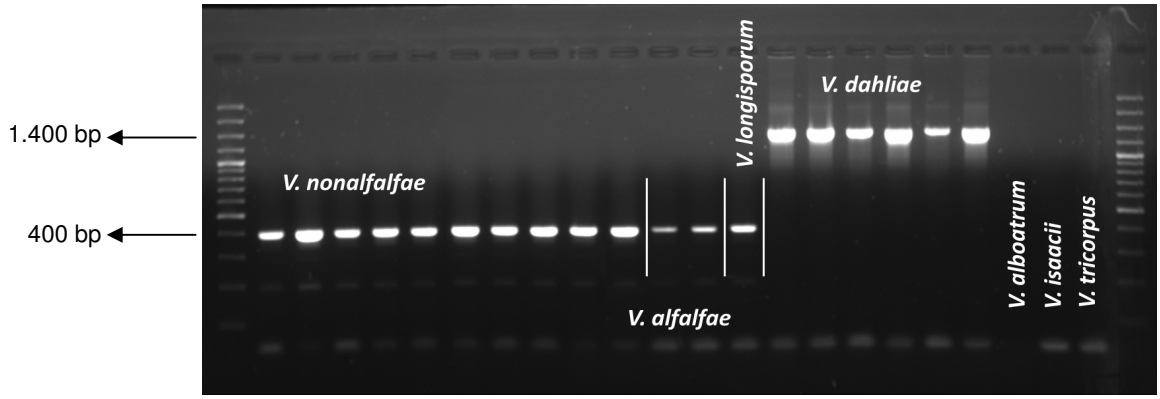


Slika 4: RT-qPCR analiza *orf414*
Analiza izražanja potencialne dolge ne-kodirajoče RN*A orf414* v dveh različnih sevih *V. nonalfalfa*e (Rec=blag, T2=letalen). Vrednosti izražanja so bile normalizirane z nivojem izražanja ubikvitina *(ub*q) kot reference. Vertikalni črti predstavljata standardni napaki treh bioloških ponovitev.

Figure 4: RT-qPCR analysis of *orf414*
Expression analysis of th*e* potential long non-coding RN*A orf414* in two different strains of *V. nonalfalfa*e (Rec=mild, T2=lethal). Expression values were normalized with ubiquitin *(ub*q) as a housekeeping reference. Bars indicate standard errors of three biological replicates.

4.1.4 Analiza dolžinskega polimorfizma

Med poravnavo *V. nonalfalfae* in *V. dahliae* mitohondrijskih genomov smo odkrili 1.221 bp veliko regijo, ki je bila prisotna med genom *coxI* in tRNA za prolin samo pri *V. dahliae*. S kartiranjem *V. nonalfalfae* mitohondrijskih odčitkov na mitohondrijski genom *V. dahliae* smo to najdbo še dodatno potrdili (Priloga F). To regijo smo nato preverili še s PCR pomnožitvijo v 22 različnih *Verticillium spp.* izolatih (Preglednica 11). Začetne oligonukleotide smo pripravili z ozirom, da so premeščali dolžinski polimorfizem preko ohranjenih regij v obeh glivah, s pričakovano dolžino amplikona 1.400 bp (*V. dahliae*) in 400 bp (*V. nonalfalfae*). Rezultati PCR pomnoževanja so vidni na spodnji sliki (Slika 5).



Slika 5: Primer pomnoževanja mitohondrijskega polimorfizma pri 22 sevih rodu *Verticillium*
Figure 5: A sample PCR amplification of the mitochondrial polymorphism in 22 strains of the *Verticillium* lineage

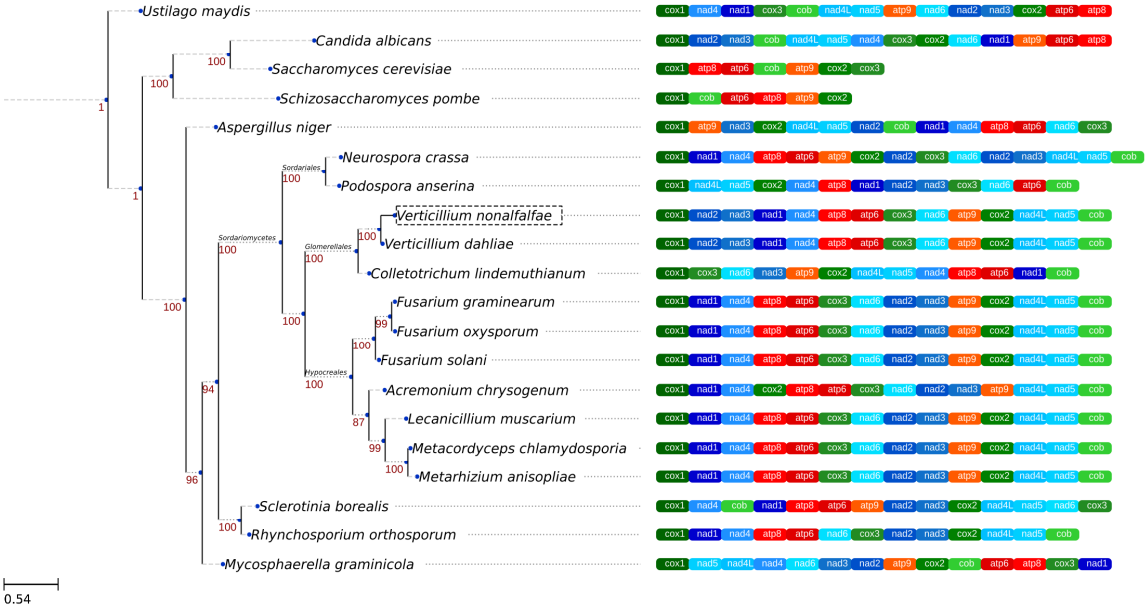
Preglednica 11: Obravnavani sevi rodu *Verticillium* pri analizi mitohondrijskega polimorfizma
Table 11: *Verticillium* lineage strains evaluated at the mitochondrial polymorphism analysis

Organizem	Sevi
V. nonalfalfae	P10, P15, P83, T2, T6, 1974, 1953, CBS 321.91, AR0/140, 340646
V. alfalfae	41, KANADA11
V. longisporum	CBS 110.218
V. dahliae	CIG3, JKG2, PAP, V 138 I, PD335, PD584
V. alboatrum	PD639
V. isaacii	JKG20
V. tricorpus	CBS 227.84

Kasneje smo ta isti PCR postopek uporabili tudi na razširjeni zbirki 96 različnih *Verticillium spp.* izolatov (Priloga A). Pomnožitev regije je bila uspešna v 22 od 26 *V. dahliae* izolatov z daljšim amplikonom in 44 od 49 *V. nonalfalfae* s krajšim amplikonom. Daljši amplikon se je pojavil tudi pri vrsti *V. nubilum* in 400 bp amplikon se je pomnožil tudi pri *V. alfalfae* in *V. longisporum.* Pomnoževanje kateregakoli izmed amplikonov ni bilo uspešno pri vrstah *V. albo-atrum*, *V. isaacii*, *V. nigrescens* in *V. tricorpus* (Priloga G).

4.1.5 Filogenetska analiza mitohondrija z drugimi glivnimi vrstami

Filogenetsko drevo (Slika 6) prikazuje visoko bootstrap podporo za posamezne skupine. Ker imajo mitohondriji preučevanih organizmov strukturo krožne molekule, se je začetna pozicija za poravnavo arbitrarno izbrala kot *cox1* gen, ki je prisoten pri vseh vrstah v tej raziskavi.



Slika 6: Filogenetsko drevo največje verjetnosti 20 mitohondrijskih genomov, osnovano na ohranjeni skupini proteinov

Figure 6: The Maximum-likelihood phylogenetic tree of 20 mitochondrial genoms, based on a group of conserved proteins

Drevo je pripravljeno iz poravnave aminokislinskih zaporedij ohranjenih mitohondrijskih proteinov s 3.093 unikatnimi poravnanimi pozicijami in 100 bootstrap ponovitvami. Številke nad drevesnimi ločnicami predstavljajo podporne bootstrap vrednosti. Zraven drevesa je grafični prikaz poravnave mitohondrijskih protein-kodirajočih genov in njihovega zaporedja v obravnavanih vrstah. Ohranjeni nabor 14 protein-kodirajočih genov je prisoten v večini vrst, z izjemo kvasovk *S. cerevisiae* in *S. pombe*, katerim manjka družina genov NADH dehidrogenaz, *C. lindemuthianum,* kateremu manjka *nad2* gen, *N. crassa*, ki ima 2 kopiji *nad2* gena ter *R. orthosporum* in *P. anserina,* katerima manjka *atp9* gen.

Na filogenetskem drevesu so vidne 3 skupine, ki ustrezajo glivnim taksonomskim skupinam *Glomerellales, Hypocreales* in *Sordariales* znotraj družine *Sordariomycetes*

ter gruča, v katero spadata *S. borealis* in *R. orthosporum,* ki ustreza skupini *Helotiales*. Kvasovke (*S. cerevisiae, S. pombe, C. albicans*) so se združile v eno skupino in *U. maydis* ima vlogo izhodne (*angl. outgroup*) skupine. *A. chrysogenum* iz skupine *hypocreales* vsebuje translokacijo *cox2* gena, na podlagi česar se razlikuje od preostalih članov svoje skupine. Med skupinama *Glomerellales* in *hypocreales* je vidna visoka stopnja sintenije, pri čemer je edina razlika med njima translokacija *nad2-nad3* genske skupine. *V. nonalfalfae* se je pozicioniral v skupini *Glomerellales*, skupaj s *C. lindemuthianum* in bližnje sorodno *V. dahliae. C. lindemuthianum* ima precej drugačen vrstni red genov v primerjavi z ostalima članoma v skupini.

4.2 GENOMSKE ANALIZE

Veliko število genomskih analiz temelji na preiskovanju podzaporedij in njihovih lastnosti v genomu. Osnovne informacije teh podzaporedij kot so variante, ponovitvene regije, kodirajoča zaporedja in druge, so uporabne za nadaljnje analize, kjer se različne informacije združijo in iz njih pridobijo dodatni poglobljeni vpogledi v genom. Pri naši analizi genoma *V. nonalfalfae* smo najprej določili variante in jih nato združili z informacijami o kodirajočih ter ponovitvenih regijah, da smo lahko preučili vpliv variant na kodirajoče regije, določili gostoto porazdelitev genomskih značilnosti v genomu (variant, eksonov in ponovitev) in nazadnje še opravili Ka/Ks analizo, s katero smo poskusili ovrednotiti vpliv evolucijskega pritiska na kodirajoče regije genoma *V. nonalfalfae*.

4.2.1 Določevanje variant

Variante v genomu smo določili z uporabo programskega cevovoda, ki je po procesu preverjanja kakovosti odčitkov (Priloga I) in njihovega kartiranja (podrobnosti v preglednici 12) implementiral njihovo napovedovanje, filtriranje glede na minimalno in maksimalno globino pokritosti ter filtriranje glede na minimalno kakovost kartiranja. Napovedane variante smo nato še filtrirali glede na zigotnost (*V. nonalfalfae* je homozigoten organizem) in s tem določili končnih 1.337 variant. Podatke končnih variant smo nato združili še z informacijami o lokacijah kodirajočih regij v genomu ter jim na podlagi tega določili njihov vpliv na genom. Rezultati teh napovedi so predstavljeni v spodnjih Preglednicah 13, 14 in 15.

Preglednica 12: Podrobnosti kartiranja odčitkov na genome posameznih sevov
Table 12: Genomic read mapping information per individual strain

Sev	Kartirani odčitki	Nekartirani odčitki	% kartiranja	Vsi odčitki
1953	13.049.140	535.375	96,06 %	13.584.515
1985	22.439.363	389.724	98,29 %	22.829.087
P15	10.273.544	210.643	97,99 %	10.484.187
P55	9.440.626	137.903	98,56 %	9.578.529
Rec	21.378.094	547.837	97,50 %	21.925.931
T2	93.174.280	5.875.319	94,07 %	99.049.599

Preglednica 13: Variante po kromosomih *V. nonalfalfae*
Table 13: *V. nonalfalfae* variants per chromosome

Kromosom	**Dolžina (bp)**	**Število sprememb**
1	5.428.263	167
2	5.191.599	119
3	4.104.733	106
4	4.081.529	112
5	3.603.980	83
6	3.100.896	70
7	3.109.290	95
8	3.118.363	75
9	2.373.895	60
10	1.529.102	24
neuvrščen	2.940.961	426
skupaj		1.337

Preglednica 14: Število sprememb po njihovem tipu
Table 14: Number of changes according to type

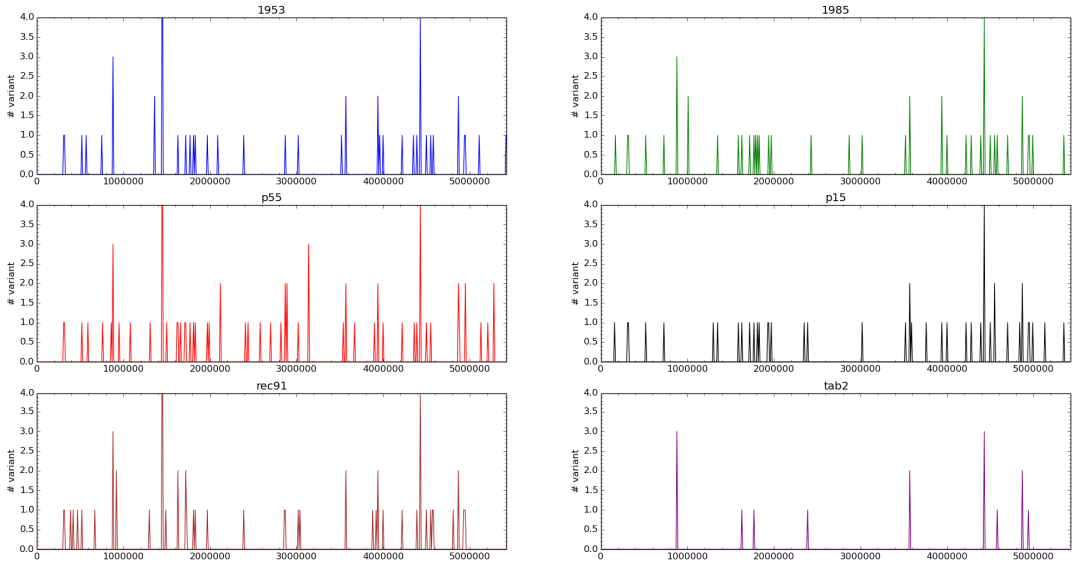
Tip Spremembe	Število sprememb
SNP	1.080
MNP	0
INSERCIJA	101
DELECIJA	152
MEŠANO	4
INTERVAL	0
SKUPNO	1.337

Preglednica 15: Število tranzicij in transverzij
Table 15: Number of transitions and transversions

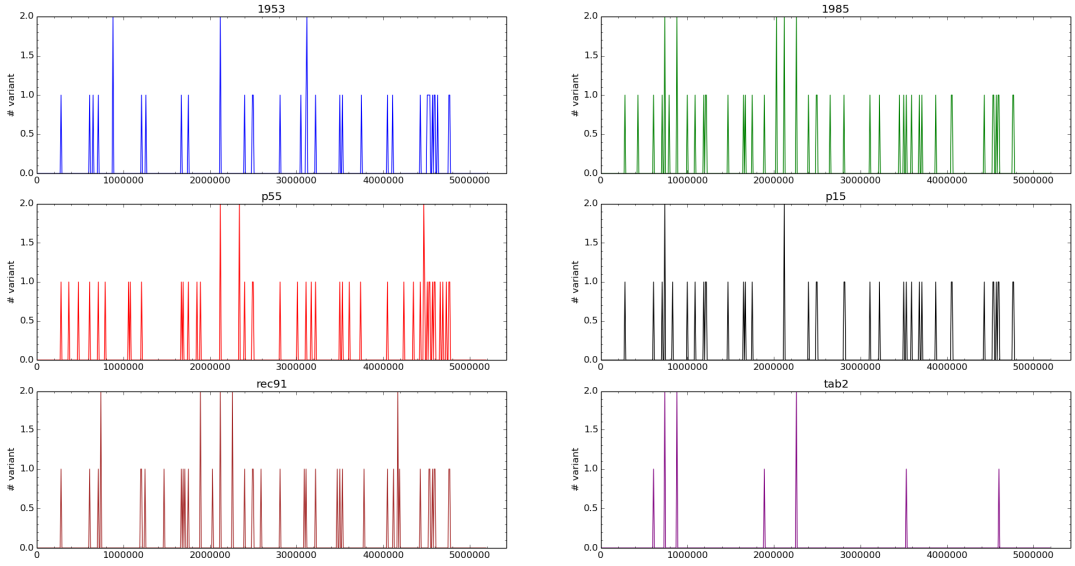
Tranzicije	Transverzije	Ts/Tv razmerje
476	258	1,845

4.2.2 Gostotne porazdelitve

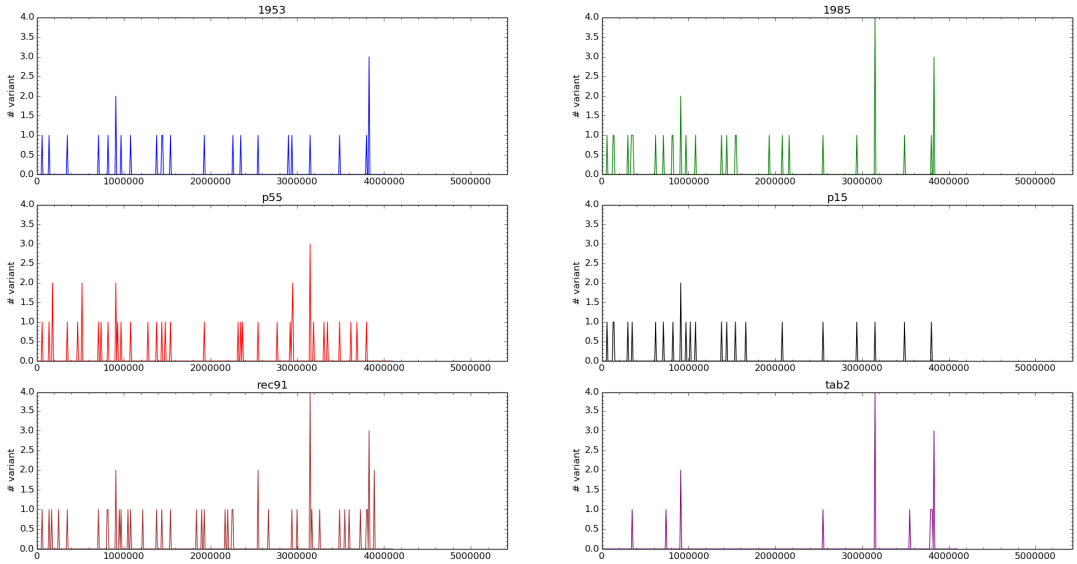
Vizualni pregled genomskih značilnosti smo naredili s pomočjo analize gostotne porazdelitve z drsnim oknom velikosti 10 kbp. Pri tem smo analizirali porazdelitve snp-jev v šestih obravnavanih sevih in eksonov ter ponovljivih regij, ki so bile določene v predhodni raziskavi na referenčnem *V. nonalfalfae* sevu. V naslednjem razdelku so predstavljeni rezultati analiz za 10 predvidenih jedrnih kromosomov in dodaten neuvrščen kromosom, ki je sestavljen iz konkateniranih sosesk, katere se v procesu združevanja sosesk niso uvrstile k nobenemu obstoječemu kromosomu. Pri analizi gostote SNP je dodatno obravnavan tudi mitohondrijski genom.



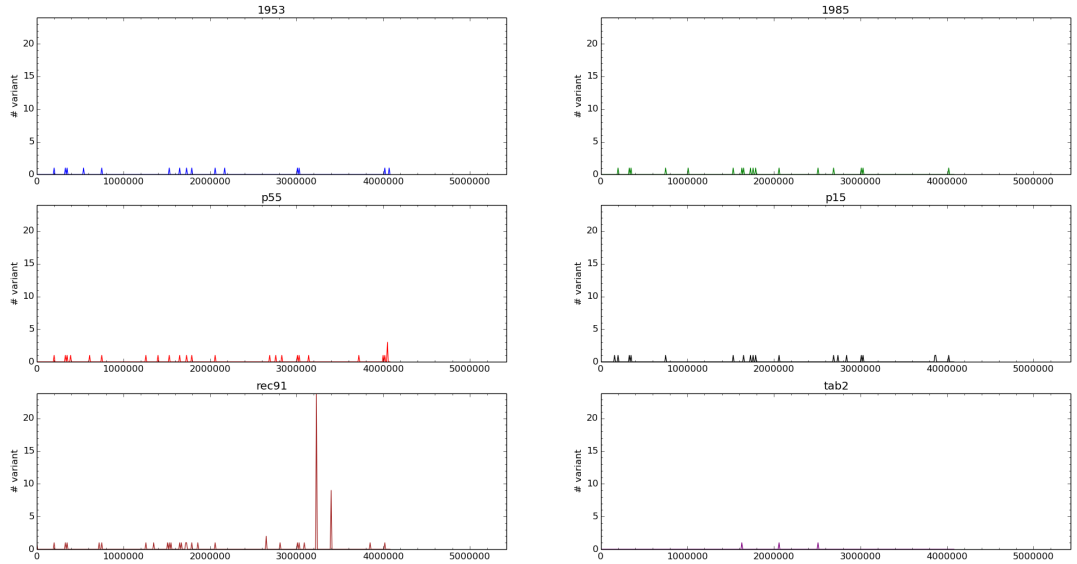
Slika 7: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 1
Figure 7: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 1



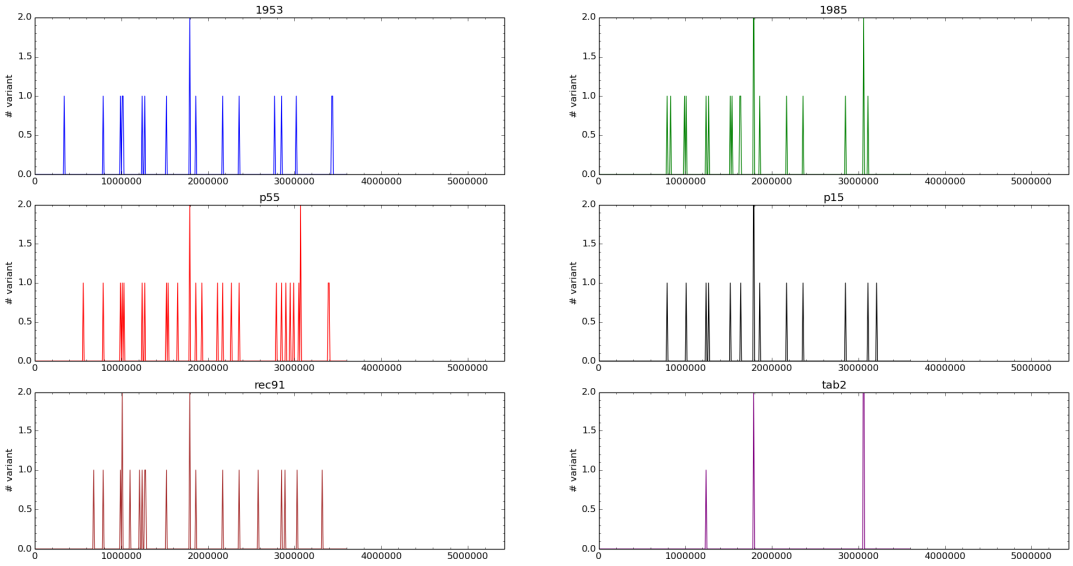
Slika 8: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 2
Figure 8: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 2



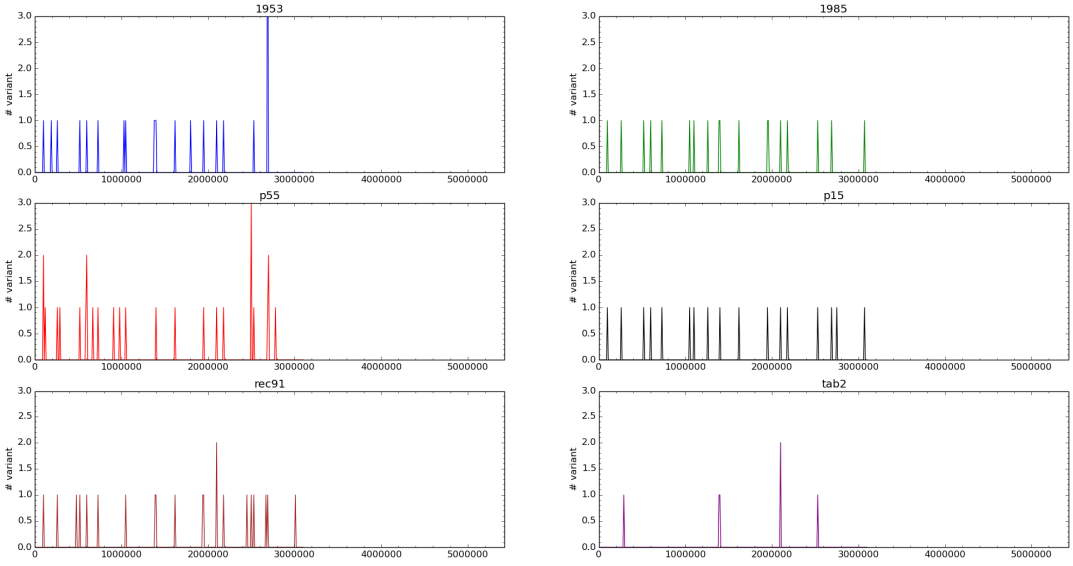
Slika 9: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 3
Figure 9: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 3



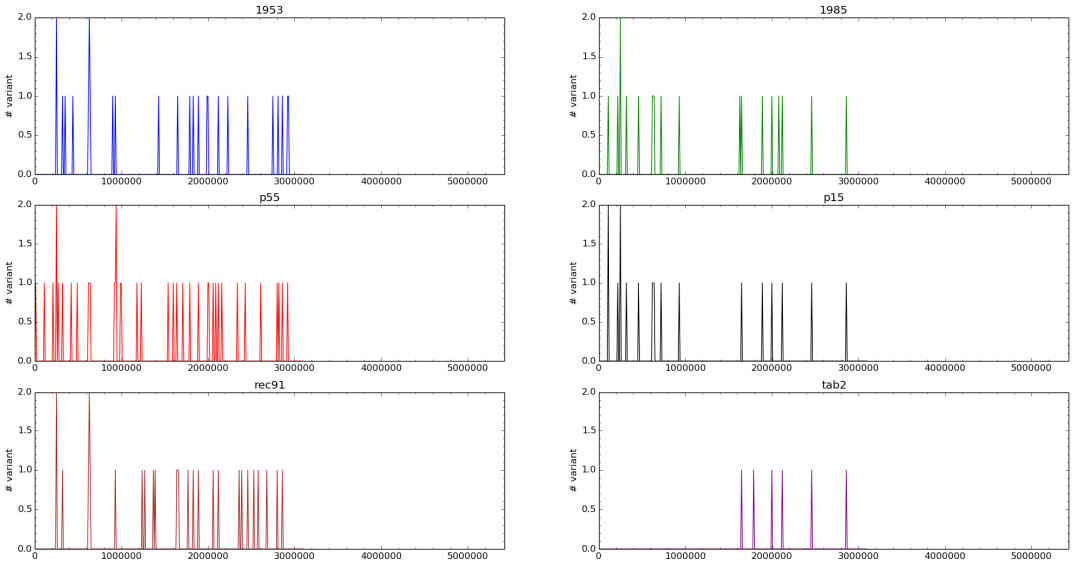
Slika 10: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 4
Figure 10: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 4



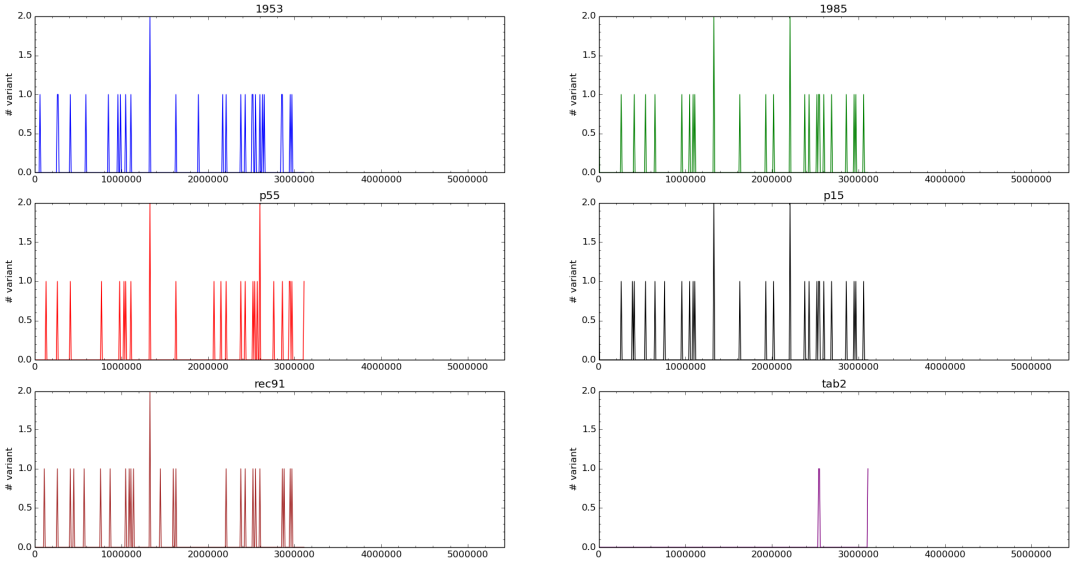
Slika 11: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 5
Figure 11: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 5



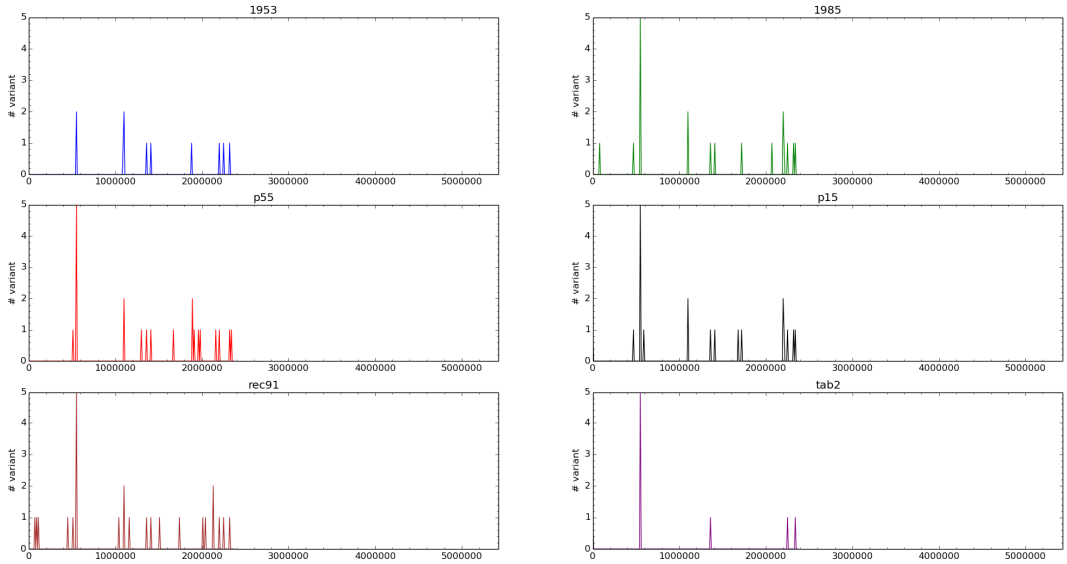
Slika 12: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 6
Figure 12: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 6



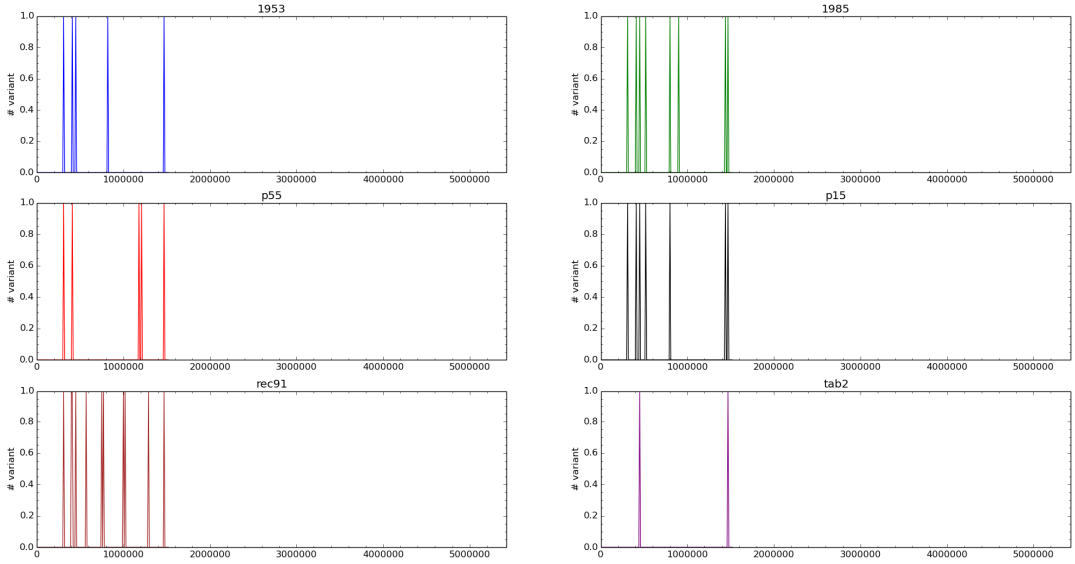
Slika 13: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 7
Figure 13: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 7



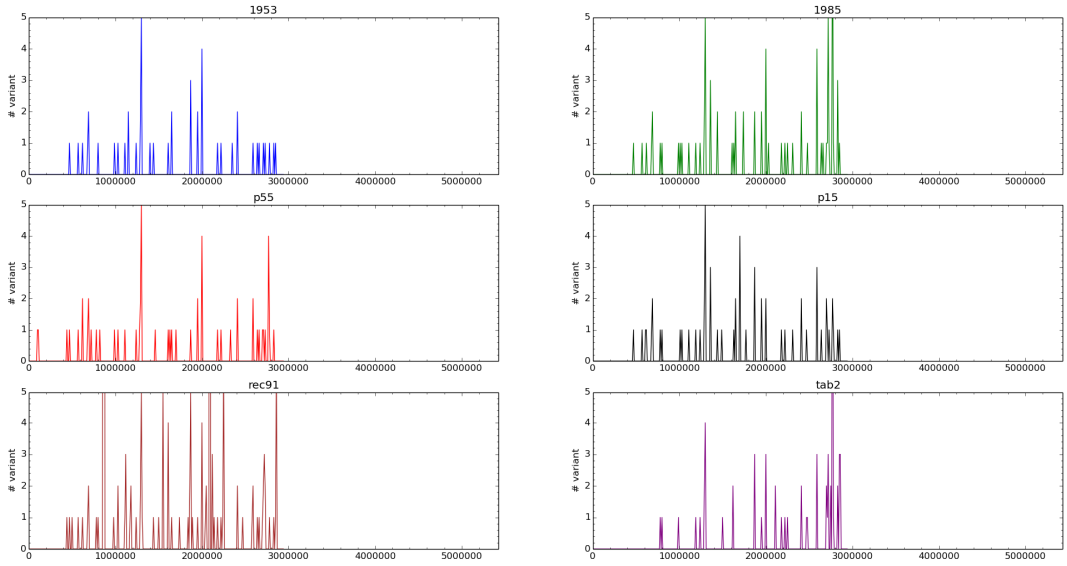
Slika 14: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 8
Figure 14: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 8



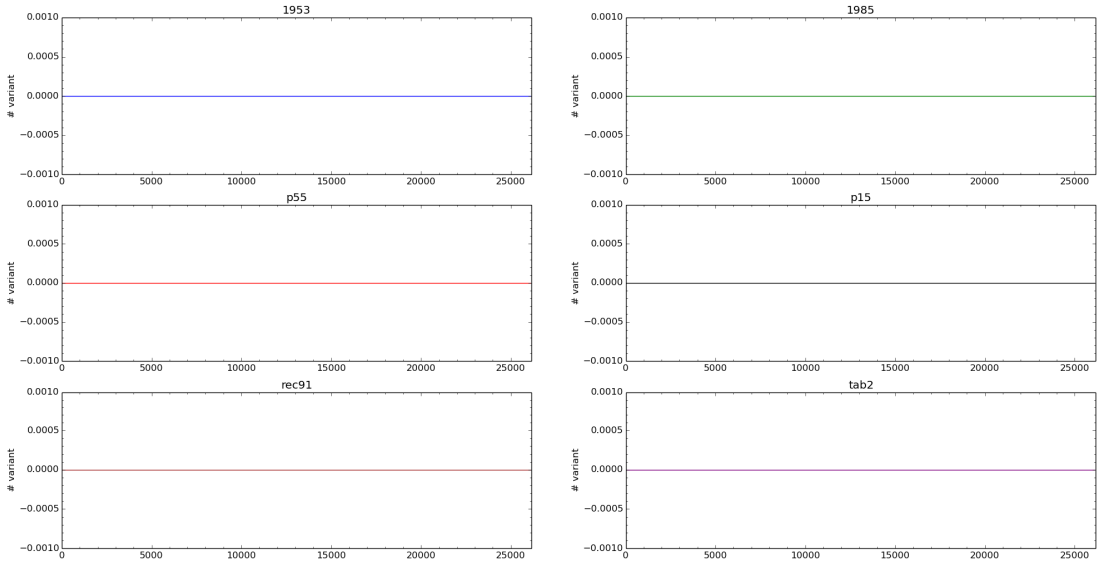
Slika 15: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 9
Figure 15: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 9



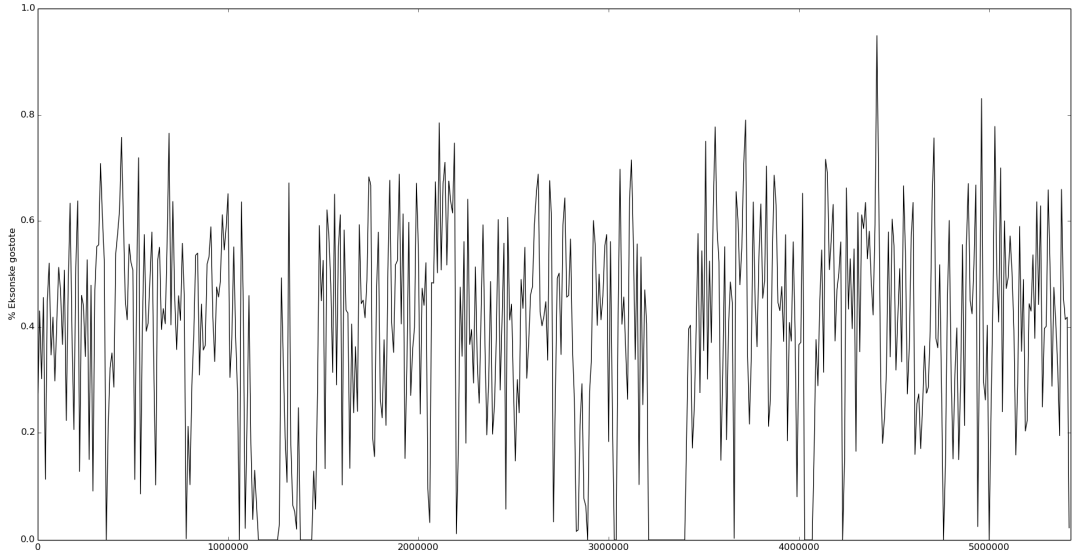
Slika 16: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za kromosom 10
Figure 16: Variant density distributions of the 6 *V. non-alfalfae* strains for chromosome 10



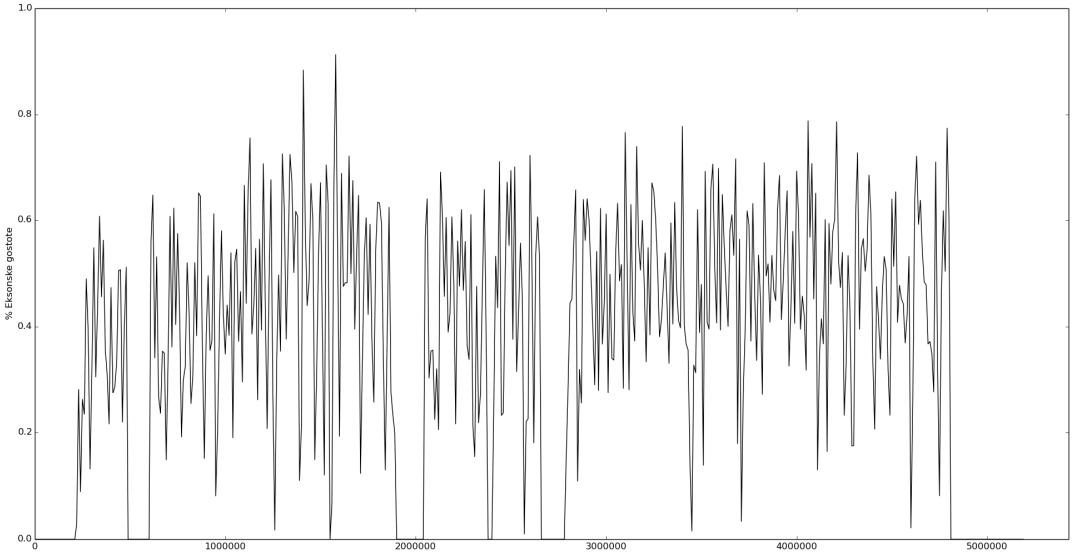
Slika 17: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za neuvrščen kromosom
Figure 17: Variant density distributions of the 6 *V. non-alfalfae* strains for the unplaced chromosome



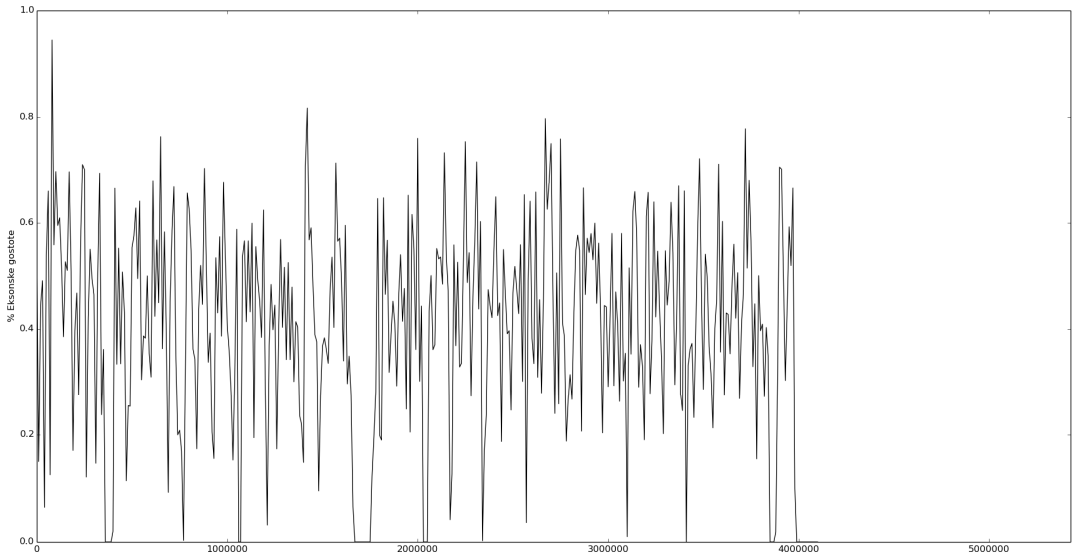
Slika 18: Gostotna porazdelitev variant za 6 sevov *V. non-alfalfae* za mitohondrijski genom
Figure 18: Variant density distributions of the 6 *V. non-alfalfae* strains for the mitochondrial genome



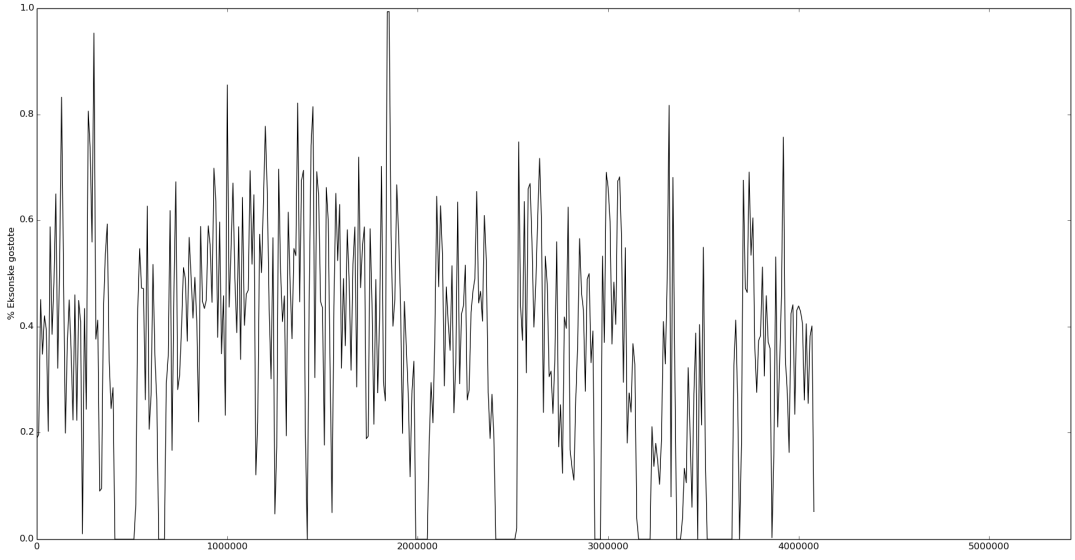
Slika 19: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 1
Figure 19: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 1



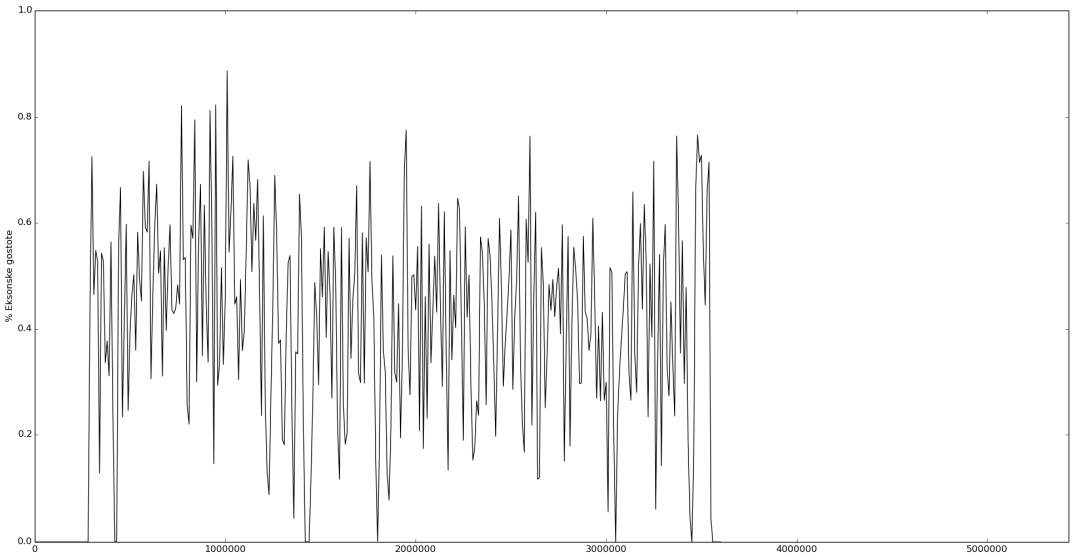
Slika 20: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 2
Figure 20: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 2



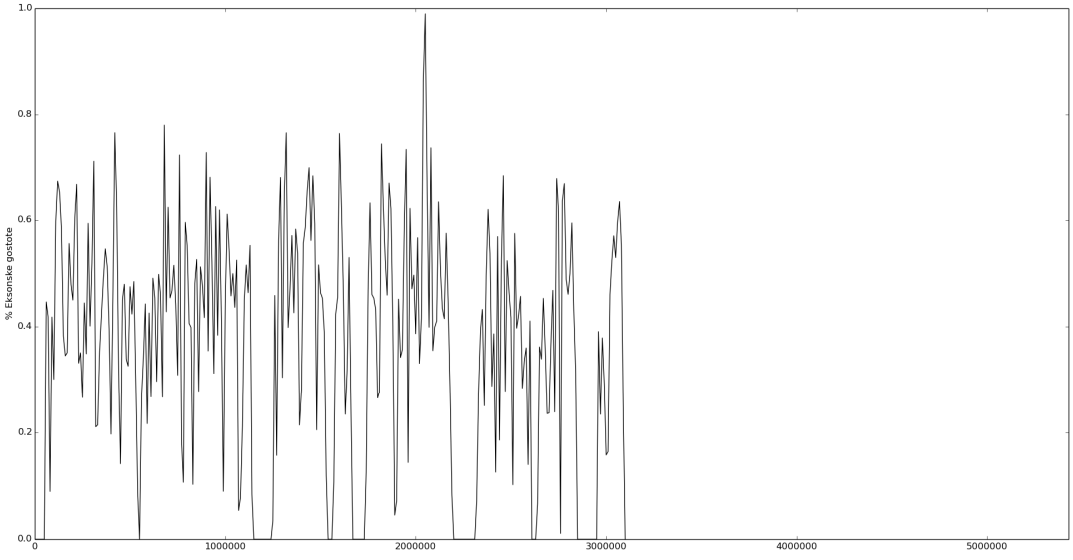
Slika 21: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 3
Figure 21: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 3



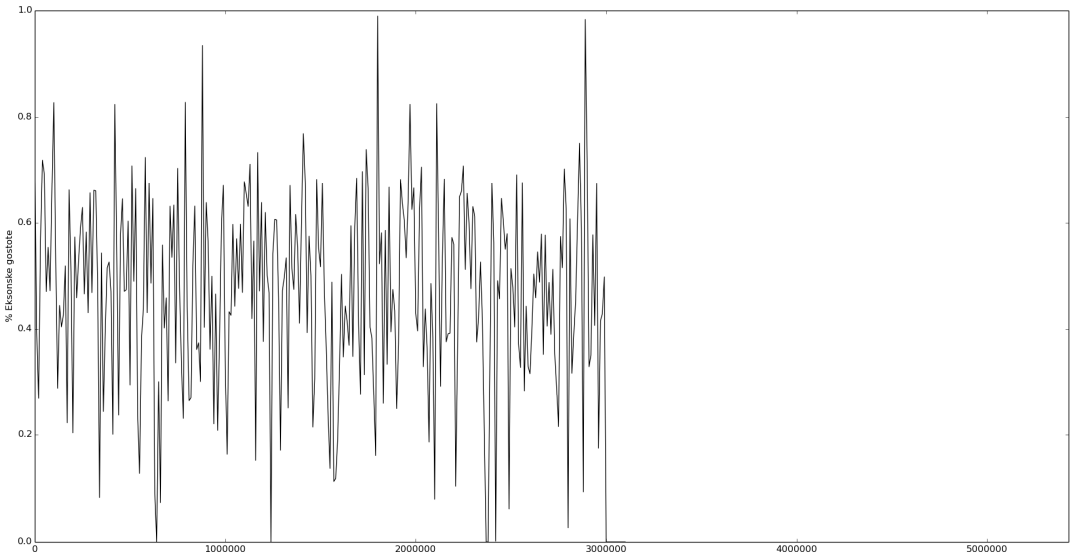
Slika 22: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 4
Figure 22: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 4



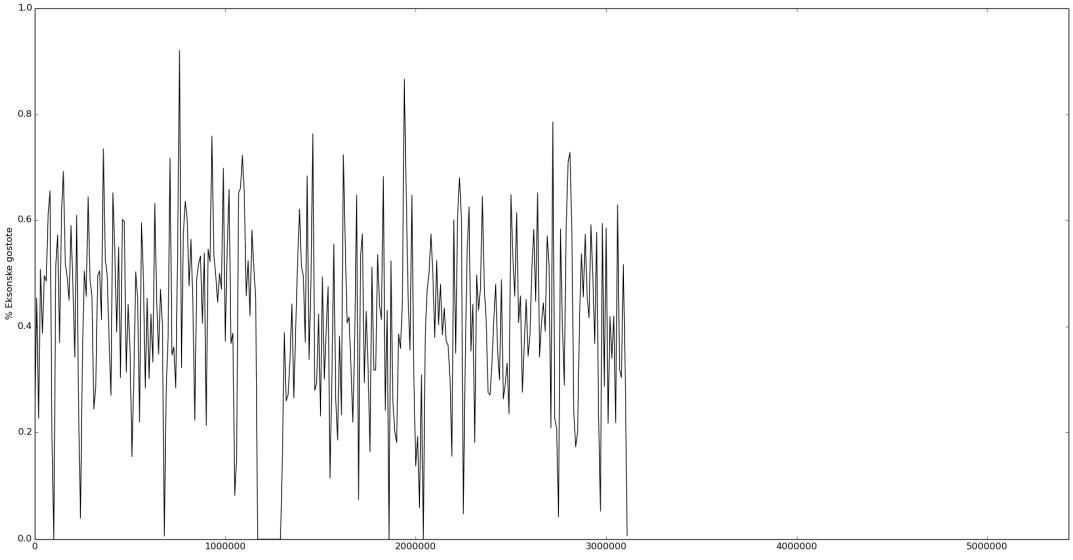
Slika 23: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 5
Figure 23: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 5



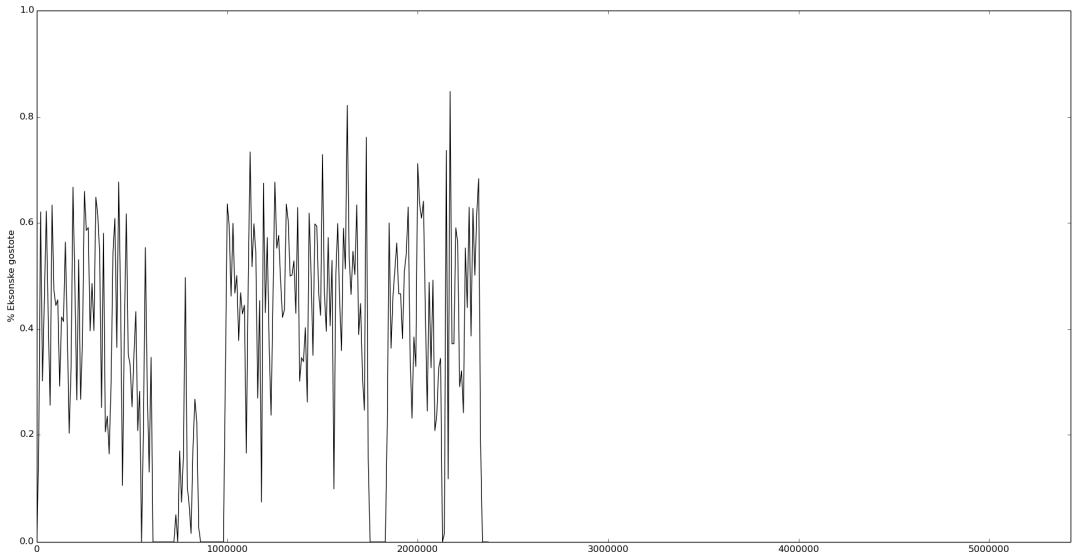
Slika 24: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 6
Figure 24: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 6



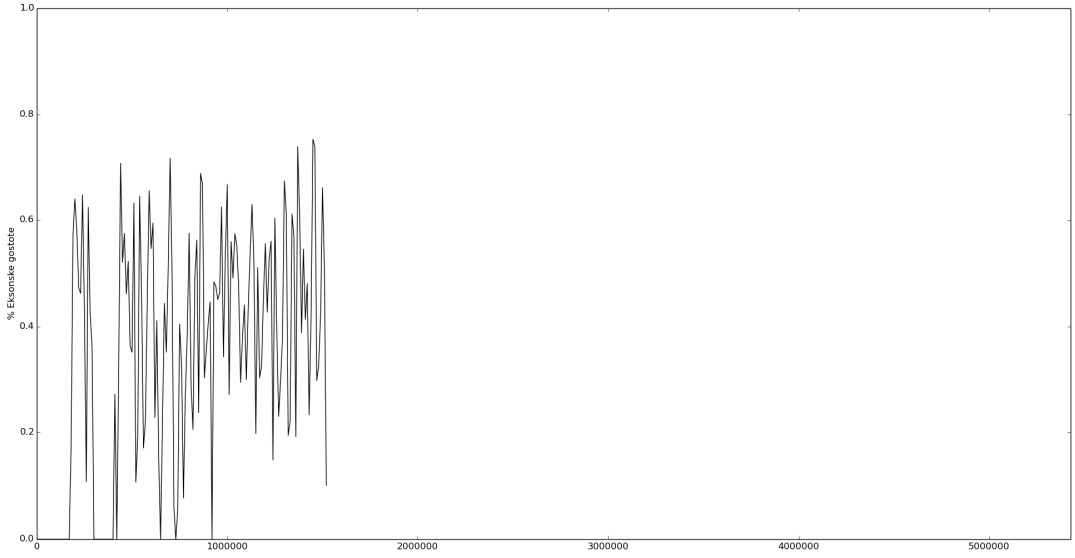
Slika 25: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 7
Figure 25: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 7



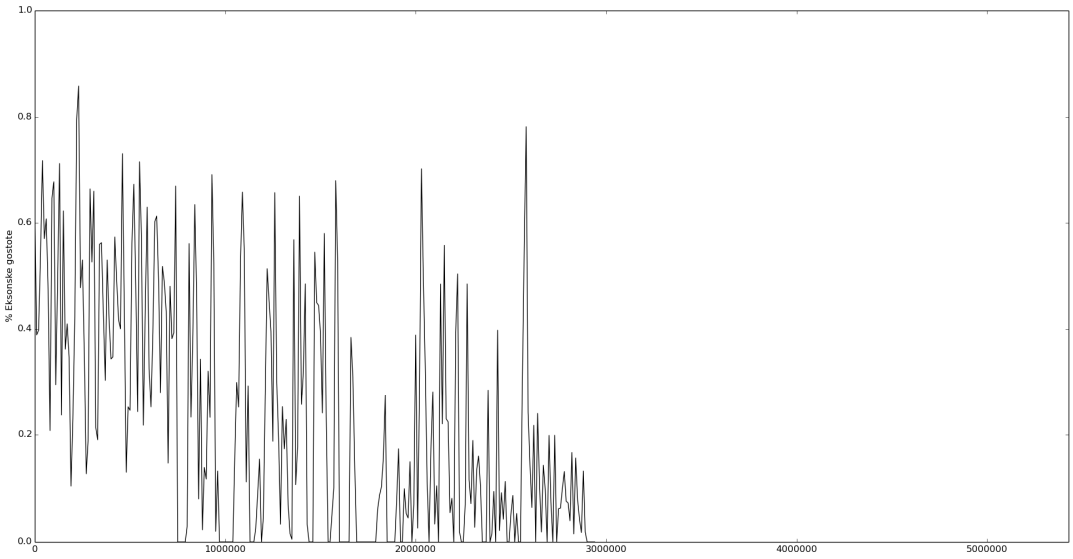
Slika 26: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 8
Figure 26: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 8



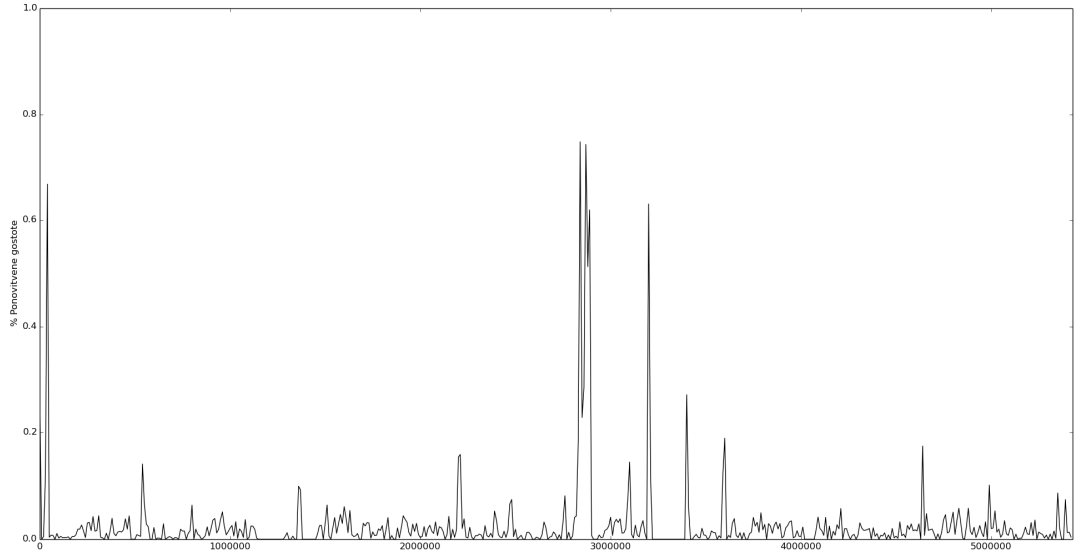
Slika 27: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 9
Figure 27: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 9



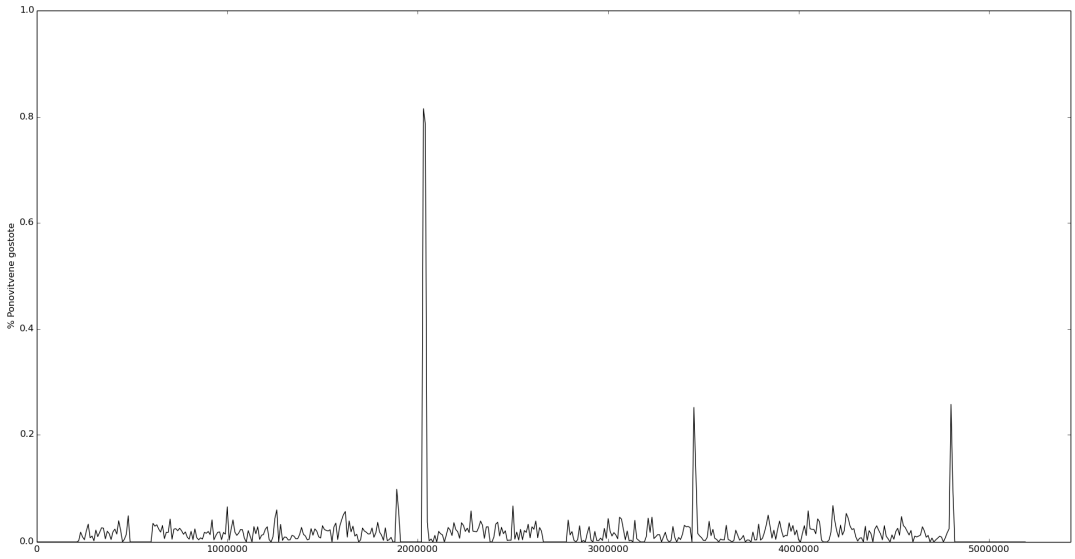
Slika 28: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za kromosom 10
Figure 28: Exon density distributions of the 6 *V. non-alfalfae* strains for chromosome 10



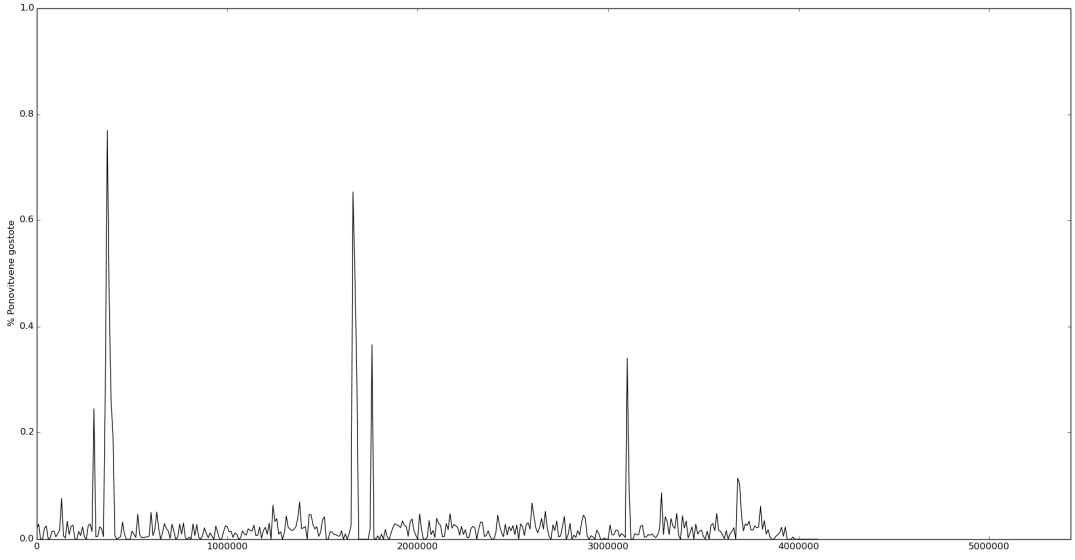
Slika 29: Gostotna porazdelitev eksonov za 6 sevov *V. non-alfalfae* za neuvrščen kromosom
Figure 29: Exon density distributions of the 6 *V. non-alfalfae* strains for the unplaced chromosome



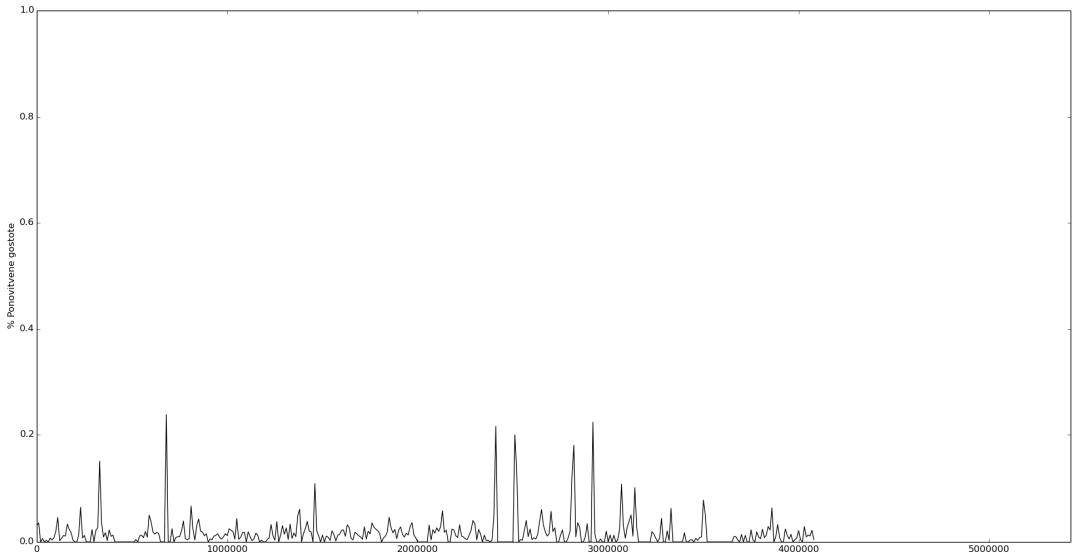
Slika 30: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 1
Figure 30: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 1



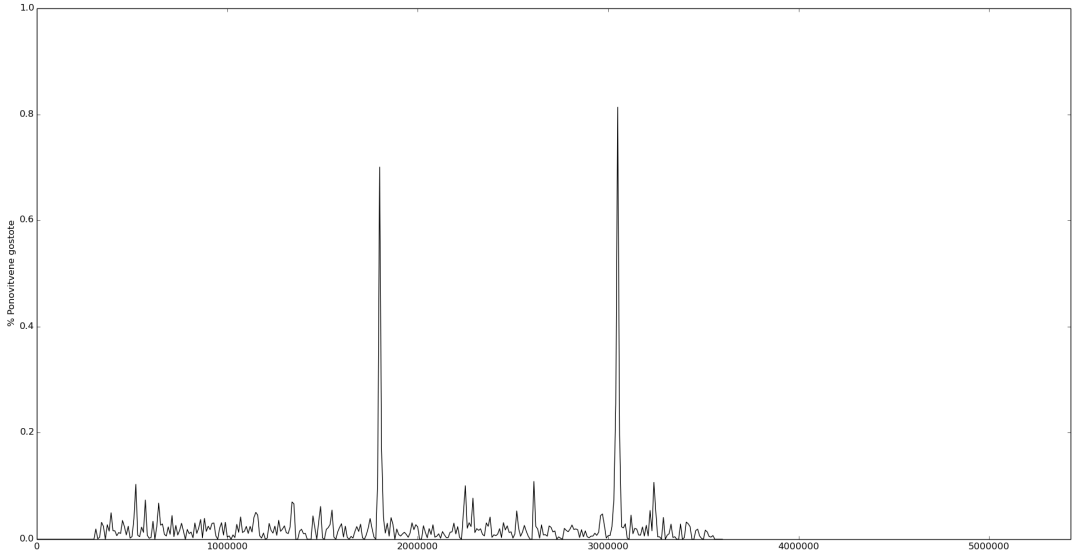
Slika 31: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 2
Figure 31: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 2



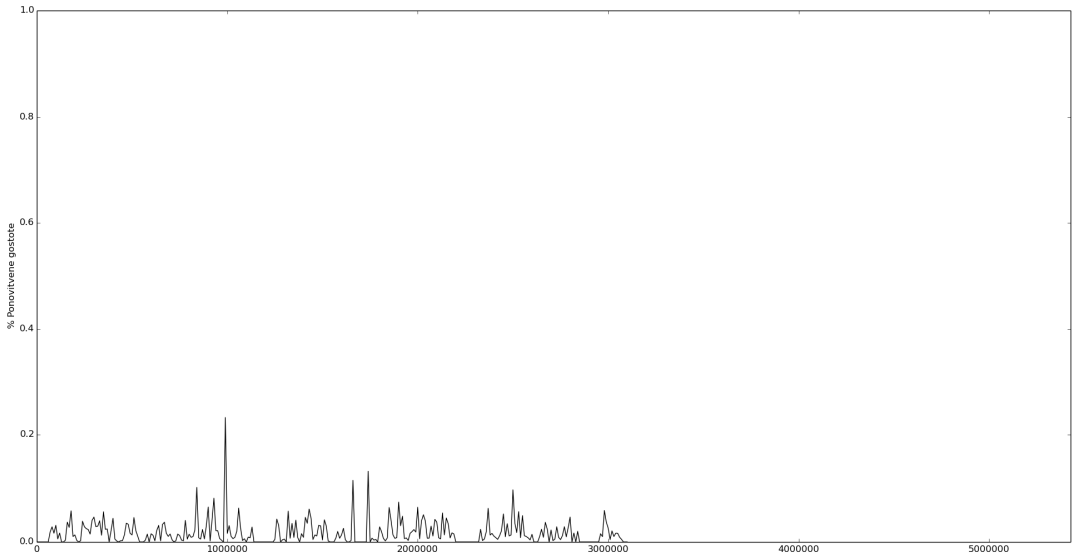
Slika 32: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 3
Figure 32: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 3



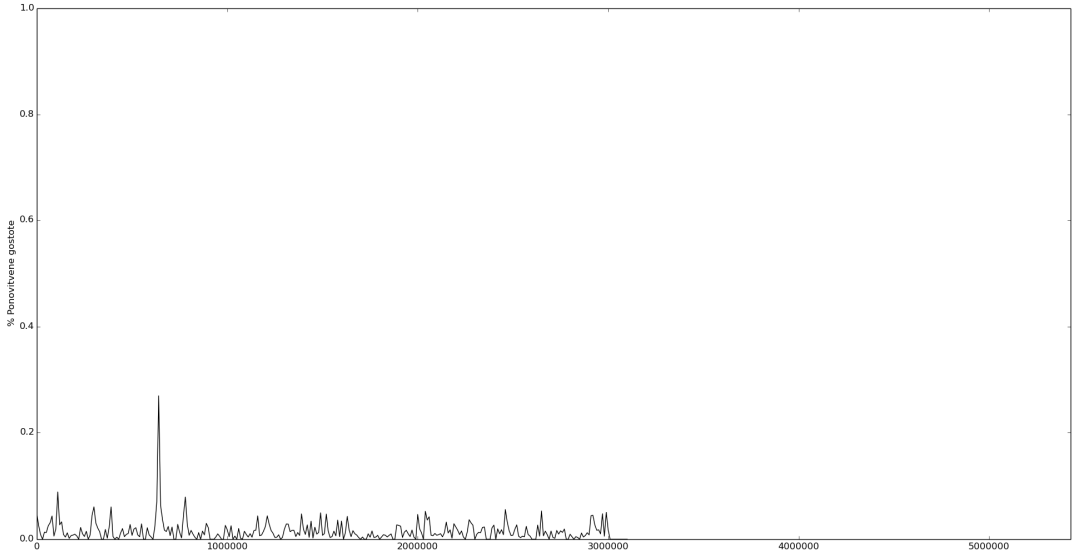
Slika 33: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 4
Figure 33: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 4



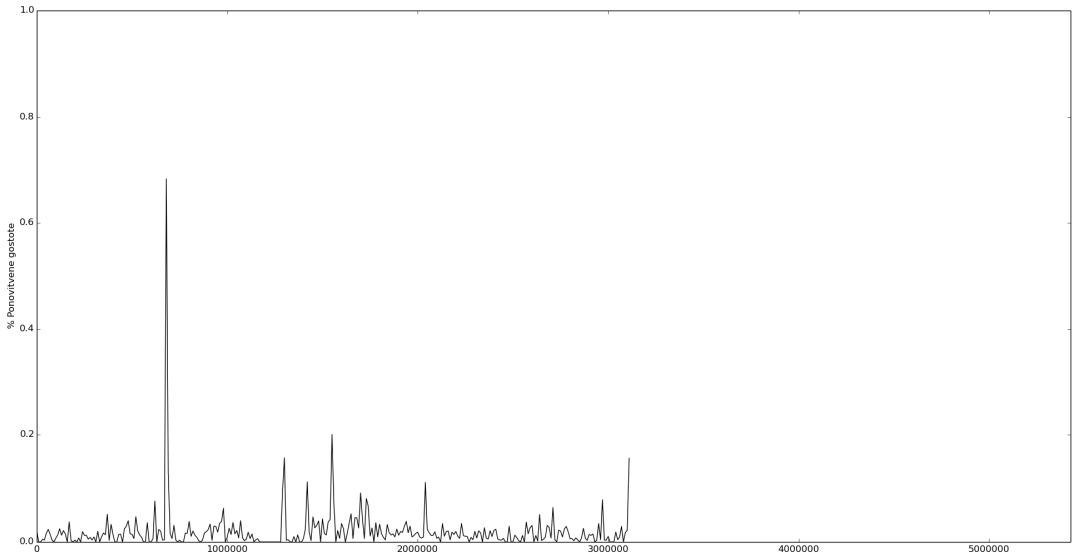
Slika 34: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 5
Figure 34: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 5



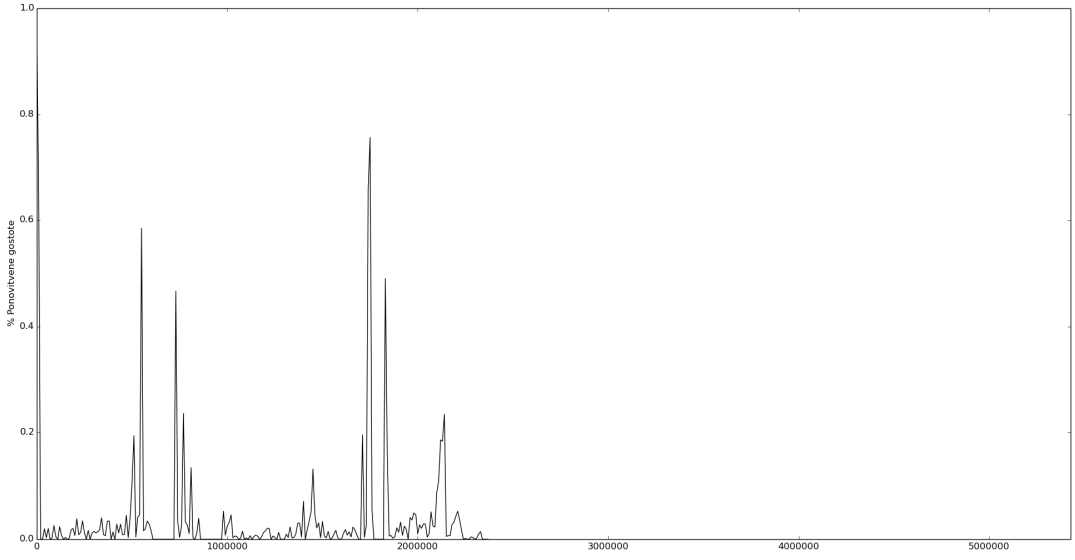
Slika 35: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 6
Figure 35: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 6



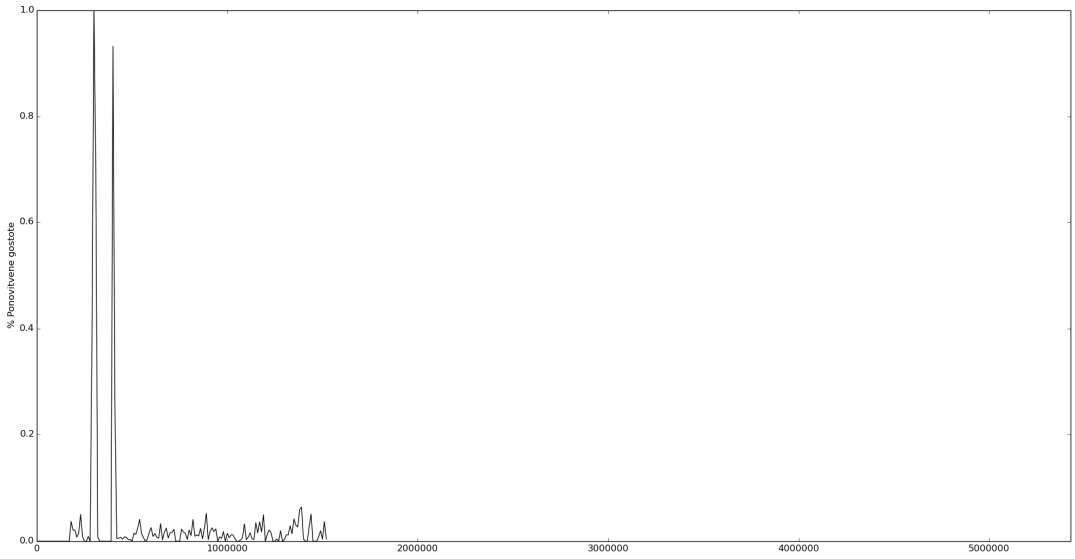
Slika 36: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 7
Figure 36: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 7



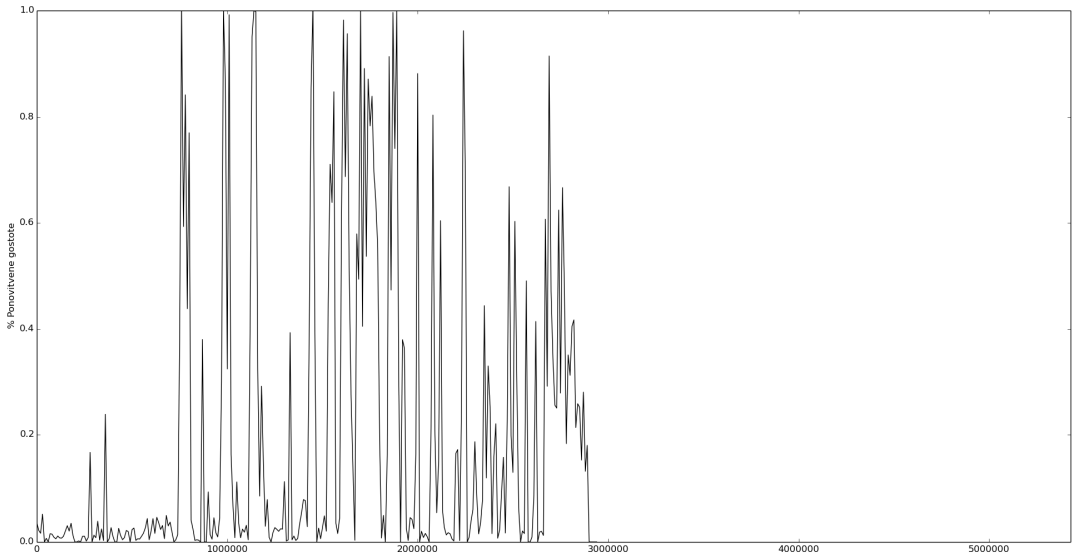
Slika 37: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 8
Figure 37: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 8



Slika 38: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 9
Figure 38: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 9



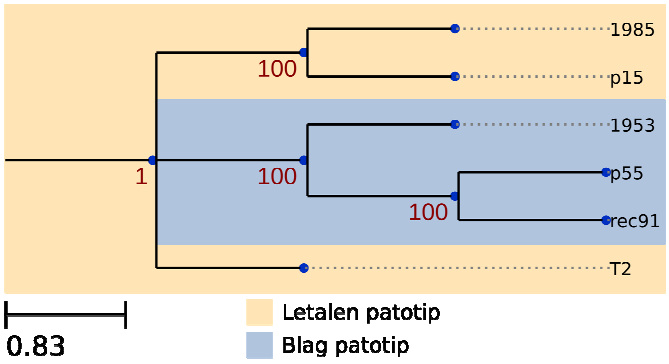
Slika 39: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za kromosom 10
Figure 39: Repeat density distributions of the 6 *V. non-alfalfae* strains for chromosome 10



Slika 40: Gostotna porazdelitev ponovitev za 6 sevov *V. non-alfalfae* za neuvrščen kromosom
Figure 40: Repeat density distributions of the 6 *V. non-alfalfae* strains for the unplaced chromosome

4.3 FILOGENETSKA ANALIZA SEVOV

Večkratna poravnava SNP variant vseh 6 sevov je imela 207 unikatno poravnanih vzorcev. RAxML obravnava vzorce kot unikatne stolpce v poravnavi, ki se pri analizi obravnavajo samo enkrat, vendar so ustrezno obteženi glede na število pojavitev v prvotnem zaporedju. Končno drevo z največjo verjetnostjo (*angl. maximum likelihood*) (vrednost -4173,0429) smo pridobili po 100 iteracijah postopka bootstrap in je prikazano na spodnji sliki (Slika 41). Vidni sta 2 skupini, ki predstavljata letalen ter blag patotip in imata obe najvišjo bootstrap podporo (100).



Slika 41: Filogenetsko drevo *V. nonalfalfae* sevov na osnovi njihovih variant
Figure 41: Variant-based phylogenetic tree of *V. nonalfalfae* strains

4.4 KA/KS

Določitev evolucijskega pritiska znotraj vr*ste V. nonalfalfae* med sevi preko Ka/Ks koeficientov nismo uspeli narediti zaradi prevelike stopnje podobnosti zaporedij. Metodološki problem je nastopil pri izračunu vrednosti Ks, saj je ta zaradi velike količine možnih sinonimnih mest in malega števila dejanskih zamenjav postala zelo majhna in so bili Ka/Ks koeficienti v skoraj vseh primerih zelo visoki. Posledično zato nismo mogli ločiti resničnih pozitivnih zadetkov od lažnih pozitivnih zadetkov.

Ta problem ne nastopi pri poravnavi bolj oddaljenih zaporedij, zaradi tega smo namesto med sevi naredili primerjavo med sorodnimi *Verticillium* vrstami. Za določitev evolucijskega pritiska smo izračunali Ka/Ks koeficiente med referenčnim genomom *V. nonalfalfae* T2 in *Verticillium* vrstama *V. alfalfae* (VaMs102) ter *V. dahliae* (VdLs17) z uporabo NG in YN metod s programom gKaks ter njihovih ortologov, določenih s programom orthoMCL. S tem pristopom smo določili 142 genov, ki so imeli Ka/Ks koeficiente večje od 1 (Preglednica 15, Priloga H).

Preglednica 16: Porazdelitev genov s Ka/Ks koeficientom večjem od 1 po kromosomih
Table 16: The distribution of genes with Ka/Ks coefficients larger than 1 along chromosomes

Kromosom	# genov s Ka/Ks > 1
1	20
2	13
3	23
4	14
5	13
6	8
7	12
8	7
9	5
10	8
neuvrščen	19

Teh 142 genov smo dodatno preverili še za obogatitev GO pojmov. Rezultati te analize obsegajo statistično značilnost obogatenost ter statistično značilno osiromašenost GO pojmov pod mejo p-vrednosti 0,05 % in so predstavljeni v spodnji preglednici 16.

Preglednica 17: Rezultati obogatitvene analize GO genov pod vplivom pozitivne selekcije
Table 17: Results of the GO enrichment analysis of genes under positive selection

GO ID	Kategorija	obogatenost	GO Pojem	p-vrednost
GO:0030150	Biološki proces	+	vnos proteinov v mitohondrijski matriks	0,0021546871
GO:0055085	Biološki proces	-	transmembranski transport	0,0026129104
GO:0043581	Biološki proces	-	razvoj micelija	0,0073089041
GO:0055114	Biološki proces	-	oksidacijsko-redukcijski proces	0,0149287385
GO:0018101	Biološki proces	+	citrulinacija proteinov	0,0271012416
GO:0048280	Biološki proces	+	fuzija veziklov z Golgijevem apartom	0,0271012416
GO:0006906	Biološki proces	+	fuzija veziklov	0,0271012416
GO:0009081	Biološki proces	+	metabolni proces razvejanih aminokislin	0,0271012416
GO:0046949	Biološki proces	+	biosintetični proces maščobne-acil-CoA	0,0271012416
GO:0019310	Biološki proces	+	katabolni proces inozitola	0,0271012416
GO:0016021	Celična komponenta	-	integralna komponenta membrane	0,0025925069
GO:0005744	Celična komponenta	+	kompleks mitohondrijske membranske presekvenčne translokaze	0,0102072576
GO:0030286	Celična komponenta	+	dineinski kompleks	0,0271012416
GO:0030659	Celična komponenta	+	citoplazemska vezikelska membrana	0,0271012416
GO:0005871	Celična komponenta	+	kinezinski kompleks	0,0271012416
GO:0034045	Celična komponenta	+	pre-avtofagosomalna strukturna membrana	0,0271012416
GO:0031201	Celična komponenta	+	SNARE kompleks	0,0271012416
GO:0030532	Celična komponenta	+	majhen jedrni ribonukleoproteinski kompleks	0,0271012416
GO:0004668	Molekularna funkcija	+	protein-arginin deiminazna aktivnost	0,0271012416

se nadaljuje

nadaljevanje preglednice 16

GO ID	Kategorija	obogatenost	GO Pojem	p-vrednost
GO:0043754	Molekularna funkcija	+	dihidrolipoillizin-ostanek (2-metilpropanoil) transferazna aktivnost	0,0271012416
GO:0004354	Molekularna funkcija	+	glutamat dehidrogenazna (NADP+) aktivnost	0,0271012416
GO:0004558	Molekularna funkcija	+	alfa-14-glukozidazna aktivnost	0,0271012416
GO:0003777	Molekularna funkcija	+	mikrotubul motorična aktivnost	0,0403596784

5 RAZPRAVA

V tej doktorski nalogi smo preučevali fitopatogeno glivo *Verticillium nonalfalfae* z vidika analize njenega mitohondrijskega genoma ter znakov pozitivne evolucije v jedrnih genih. Pri tem smo se v veliki meri poglobili na različna področja bioinformatike, ki so posamično reševala računsko in algoritmično zahtevne probleme, ob združitvi pa omogočala sintezo podatkov. V prvem delu smo z uporabo genomskih podatkov sestavili in pripisali genomske značilnosti *V. nonalfalfae* mitohondrijskemu genomu ter na podlagi tega izvedli filogenetsko analizo s sorodnimi glivni vrstami, da smo ugotovili, v katero taksonomsko skupino se najverjetneje uvrsti naš referenčni sev T2. Poleg analize mitohondrijskega genoma smo izvedli tudi mutacijsko analizo jedrnih genov, s katero smo želeli preveriti, če se katere podskupine genov razvijajo značilno drugače kot ostale in na podlagi tega osnovati morebitne evolucijske scenarije. Delo je temeljilo na vnaprej pripravljenem in referenčnem genomu s pripisanimi značilnostmi, ki je nastal med potekom projekta sekvenciranja genoma *V. nonalfalfae* na katedri za genetiko, biotehnologijo, statistiko in žlahtnjenje rastlin, Oddelka za agronomijo.

Programski cevovodi, ki smo jih razvili za pripravo mitohondrijskega genoma, so vključevali sintezo večplastnih podatkov, pripis značilnosti z večimi prediktorji in sestavljanje zaporedij z namenskimi programi. Te metode so ponavadi v uporabi pri večjih genomih, vendar so navkljub relativni kompleksnosti primerni tudi za uporabo na organelnih genomih in glede na naše izkušnje in testiranja prinašajo večjo verodostojnost v primerjavi z rezultati preprostejših postopkov, brez integracij podatkov ter kompleksnejših postopkov pripisa genomskih značilnosti. Pri snovanju cevovodov smo se poslužili odprtokodnih orodij in jih prevedli iz programske kode v delujoče programe, katere smo namestili na lastnem strežniku. Najprej smo vzpostavili posamezne korake analiz, katere smo sčasoma združevali v daljše postopke in te na koncu strnili v programske cevovode. Z njimi smo nato lahko hitreje izvajali testiranja programov, pristopov, izbir parametrov, načine vizualizacij in prikaze rezultatov. To je omogočilo optimizacijo programskih cevovodov za naše nabore podatkov. Pri vsem tem smo se redno posluževali iterativnih izboljšav s preučevanjem dobrih praks, iskanjem relevantnih informacij, učenjem novih metod in pristopov ter na podlagi novih spoznanj stalno spreminjali in dopolnjevali oz. tudi spreminjali naše programske cevovode. Ko je število teh naraslo nad mejo preproste obvladljivosti, smo se odločili, da za njihovo lažje izvajanje vzpostavimo ogrodje (*angl. framework*), s katerim si olajšamo izvajanje preko modularne sestave in jih v tem okvirju optimiziramo. Prvi korak je bila implementacija avtomatizacije in paralelizacije izvajanja postopkov na enem strežniku, (kjer je bilo možno) hkrati pa smo poskrbeli za popolno reproducibilnost končnih rezultatov iz surovih podatkov. Zraven smo vključili še izvoz rezultatov v različnih formatih in izvedbo vizualizacij po meri, kjer je to bilo možno oz. izvedljivo.

Implementacija ogrodja, ki se izvaja v Linux ukazni vrstici, je bila napisana v programskih ter skriptnih jezikih Python, Bash in Awk. Izbor operacijskega sistema Linux so usmerjala predvsem odprtokodna orodja, ki so namenjena samo za to platformo.

Pri pridelavi hmelja so se dosedaj pojavile letalne oblike *V. nonalfalfae* na treh različnih geografskih območjih (Anglija, Nemčija, Slovenija) (Keyworth, 1942; Radišek in sod., 2003; Seefelder in sod., 2009). Hipotezo, katero smo na podlagi pojave teh sevov želeli testirati je bila, ali so se letalne oblike razvile iz prilagoditve blagih oblik na specifične gostitelje, od katerih je bil v našem primeru to hmelj. Za preverjanje predlagane hipoteze smo se odločili uporabiti več pristopov: analizo mitohondrijskih genomov izoliranih sevov, analizo filogenetskih odnosov teh sevov glede na njihove genomske variante ter analize evolucijskega pritiska na kodirajoča zaporedja v njihovih genomih preko Ka/Ks koeficientov. Poleg omenjene hipoteze smo želeli preveriti tudi, če je razvoj letalnih sevov v vsaki državi potekal neodvisno, ali pa so to le potomci enega izvornega seva.

Ko smo sestavili končne mitohondrijske kontige šestih unikatnih *V. nonalfalfae* sevov, smo jih medsebojno primerjali, da bi lahko odkrili potencialne variacije. Podoben pristop so uporabili (Jung in sod., 2012) za razlikovanje med 18 sevi *Lachancea kluyveri*, ki so bili izolirani iz različnih geografskih lokacij in ekoloških niš. Odkrili so veliko raznolikost v medgenskih regijah, ki so vsebovale variante in indele ter tudi visoko ohranjene kodirajoče regije v mitohondrijskih genomih. V našem primeru smo tudi sami odkrili povsem ohranjene kodirajoče regije, vendar v nasprotju z omenjeno raziskavo variabilnih medgenskih regij nismo zaznali. Sestavki 3 različnih programov za sestavljanje zaporedij so pokazali, da so njihova nukleotidna zaporedja identična. Ponovno kartiranje odčitkov na sestavke pa je dodatno potrdilo, da se variante na njih ne nahajajo.

Teh razlik med našo in prej omenjeno raziskavo ni mogoče enostavno določiti, sklepamo pa, da so lahko posledica razlik v velikosti mitohondrijskih genomov organizmov (*V. nonalfalfae:* 26.139 bp*, L. kluyveri:* 51.679 bp), pri čemer je večji genom statistično bolj verjetno dovzeten za mutacije. Možna je tudi razlika v drugačnih evolucijah kvasnih mitohondrijskih genomov in mitohondrijskih genomov sordariomicet ali pa je to specifična lastnost *L. kluyver.*

Naš *V. nonalfalfae* mitohondrijski genom je med najmanjšimi objavljenimi genomi z velikostjo 26.139 bp (max – *S. borealis*: 203.051 bp, min – *L. muscarium*: 24.499 bp) in ima povprečno GC vsebnost 26,92 %, kar je precej podobno GC vsebnosti

mitohondrijskega genoma *V. dahliae* (27,32 %), čeprav je še vedno precej na nizki stopnji povprečne GC vsebnosti v primerjavi z ostalimi glivnimi vrstami, ki so bile obravnavane v naši raziskavi (29,16 %, Preglednica 1). Število tRNA genov je na podobni ravni z ostalimi člani *Glomerellales* in *Hypocreales* taksonomskih skupin (Preglednica 1). Vrstni red genov za *V. nonalfalfae* mitohondrijski genom, začenši z genom *cox2* je *cox2-atp9-nad6-cox3-atp6-atp8-nad4-nad1-nad3-nad2-cox1-cob-nad5-nad4L,* kar je enak vrstni red genov kot pri *V. dahliae* mitohondrijskem genomu. Poleg ohranjene sintenije lahko vidimo prostorsko ohranjenost tudi pri RNA-kodirajočih genih, kjer imajo tRNA in rRNA skoraj enako postavitev na obeh mitohondrijskih genomih, z izjemo *V. dahliae,* kateremu manjka tRNA-Cys (GCA) in ima tako samo 25 napovedanih tRNA genov.

Večino mitohondrijskih protein-kodirajočih genov, vpletenih v oksidativno-fosforilacijski proces in sintezo ATP, je visoko ohranjenih znotraj glivnih mitohondrijskih genomov (Pantou in sod., 2006; Stone in sod., 2010; Zhao in sod., 2013; Kouvelis in sod., 2004; Cardoso in sod., 2007). To se je izkazalo tudi v primeru našega mitohondrijskega genoma, saj vsebuje celoten nabor 14 "standardnih" mitohondrijskih protein-kodirajočih genov. Večino napovedanih genov se nahaja na isti verigi mitohondrijskega genoma, z izjemo dolge rRNA podenote, ki je bila napovedana na nasprotni verigi. Kodirajoče regije, ki se povečini nahajajo na isti verigi, so pogosta značilnost askomicetnih mitohondrijskih genomov (Pantou in sod., 2006; Stone in sod., 2010; Mardanov in sod., 2014; van de Sande, 2012). Dodatna specifična lastnost za taksonomsko uvrstitev *V. nonalfalfae* mitohondrijskega genoma je gen *rps3,* ki se nahaja znotraj introna. Ta posebnost je bila omenjena kot pogosta značilnost glivnih mitohondrijskih genomov iz družine sordariomicet (Sethuraman in sod., 2009). Omenjeni *rps3* gen nosi zapis za ribosomalni protein, ki je komponenta 40S podenote, kjer tvori del domene, v kateri se prične prevajanje. Vse glivne vrste v tej raziskavi, ki so člani družine sordariomicet, vsebujejo gen *rps3* znotraj introna v njihovih mitohondrijskih genomih glede na njihove Genbank zapise (*F. graminearum, F. oxysporum, F. solani, L. muscarium, M. chlamydosporia, M. anisopliae, N. crassa, P. anserina, R. orthosporum, S. borealis, V. dahliae*). Sprva je bilo videti, glede na njune Genbank zapise, da *A. chrysogenum* manjka *rps3* gen in *C. lindemuthianum* dolga rRNA podenota, vendar smo ju z ločeno raziskavo v njunih mitohondrijskih genomih našli in potrdili (podatki niso prikazani).

Pri naši primerjavi *V. nonalfalfae* (26.139 bp) z bližje sorodno *V. dahliae* (27.184 bp) smo odkrili, da sta mitohondrijska genoma primerljivih velikosti, da imata ohranjen vrstni red genov in 98,15 % identičnost na nukleotidnem nivoju. Te indikacije za taksonomsko uvrstitev našega seva so bile dodatno podprte še s filogenetsko analizo, ki

je umestila *V. nonalfalfae* mitohondrijski genom v skupino *Glomerellales* znotraj vrste *Verticillium*, skupaj z *V. dahliae* kot najbližnjim sorodnikom. Ob izboru vrst za filogenetsko analizo smo se želeli primarno osredotočiti na patogene glive, zato so te predstavljale večino od izbranih vrst. Zraven teh smo dodali še 3 predstavnike kvasovk, ki so skupaj z ostalimi patogenimi glivami iz skupine Pezizomycotina tvorile skupino askomicet. Za izhodno skupino (*angl. outgroup*) smo izbrali bazidiomicetnega patogena *Ustilago maydis* (Preglednica 1) in tako v raziskavo skupno vključili 20 glivnih vrst. Poglavitni ključ pri izbiri teh je seveda bila kakovost pripisa značilnosti njihovega mitohondrijskega genoma, kar smo določevali s pregledom njihovih objav ter z ročnim pregledom vnosov v podatkovnih bazah.

Naše filogenetsko drevo (Slika 6) je skladno s filogenetskim drevesom iz jedrnih zaporedij z ozirom na taksonomsko skupino *Sordariomycetes* (Morgenstern in sod., 2012), čeprav so mitohondriji dostikrat podvrženi povečani stopnji mutacij v primerjavi z jedrnim genomom, večji diverziteti v vrstnem redu genov in ne-kodirajočih regijah (Al-Reedy in sod., 2012). Filogenetsko drevo na podlagi jedrnih genov (Morgenstern in sod., 2012) pokaže podobne skupine kot pri mitohondrijski filogeniji, kakor npr. *Verticillium spp.* v skupini *Glomerellales*, *Fusarium spp.* v skupini *Hypocreales* in *N. crassa* z *P. anserina* v skupini *Sordariales*. Druge skupine ne izkažejo neposredne podobnosti z genomskim filogentskim drevesom, vendar to je lahko posledica slabe resolucije, zaradi pomanjkanja taksonomskih vrst v teh predelih filogenetske analize mitohondrijskih genomov.

Še ena posebna lastnost *V. nonalfalafe* mitohondrijskega genoma je prisotnost domnevne dolge ne-kodirajoče RNA (*orf414*, Fig. 1). *orf414* je ovrednotena kot dolga ne-kodirajoča RNA ker nismo našli statistično značilnih zadetkov v nr bazi z BLASTx algoritmom, prisotnost pa je bila podprta s kartiranjem RNA-Seq podatkov in dodatnim qPCR eksperimentom, ki je pokazal njeno izražanje v obeh slovenskih patotipih. Kot dodaten dokaz smo uporabili tudi vizualizacijo RNA-Seq podatkov (po 100 bp odsekih in normaliziranih z DESeq metodo) na zaporedju mitohondrijskega genoma (Slika 3). BLASTn poravnava *orf414* na glivne genome v bazah JGI MycoCosm in NCBI je pokazala, da je zaporedje unikatno samo za mitohondrije vrste *Verticillium* (*V. dahliae*). S poravnavo zaporedja *orf414* na *V. dahliae* mitohondrijski genom smo ugotovili da se nahaja med genoma *cox3* in *nad6*, kar je isti lokus kot pri *V. nonalfalfae,* s 95 % identičnostjo zaporedja. Poravnava *orf414* na mitohondrijski genom *V. alfalfae* v WGS odseku NCBI podatkovnih baz je pokazala samo delni zadetek, s 40 % podobnostjo. Glede na predhodno raziskavo (Martin, 2010) se je ta regija že uporabljala kot orodje za razlikovanje med izolati *V. dahliae.* Tu je bila regija med genoma *cox3* in *nadh6* (naš

orf414) zaznana kot zelo polimorfna in uporabna tudi kot marker za ločevanje *V. dahliae* izolatov v podskupine.

Pri primerjavi *V. nonalfalfae* in *V. dahliae* mitohondrijskih genomov smo odkrili še eno zanimivo značilnost: dodatno regijo (1.221 bp) med genom *cox1* in tRNA, ki kodira prolin pri *V. dahliae*. PCR pomnožitev te regije je pokazala, da jo vsebujejo vsi analizirani izolati *V. dahliae*, kar je razvidno iz pomnožitve 1.400 bp pomnožka (Slika 5), pri razširjeni raziskavi pa je bila pomnožitev uspešna pri 22 od 26 izolatov *V. dahliae* in se je pojavila tudi pri enem izolatu vrste *V. nubilum* (Priloga G). Glede na rezultate bi se lahko ta regija uporabila kot potencialni marker za razločevanje med *V. dahliae* in drugimi *Verticillium* vrstami kot npr. *V. nonalfalfae*. Trenutne molekularne metode za razlikovanje med temi vrstami temeljijo na sekvenciranju ITS regije ribosomalne RNA s specifičnimi PCR začetnimi oligonukleotidi (EPPO Bulletin, 2007). Prednost tega predlaganega molekularnega markerja je veliko število kopij molekul mitohondrija zaradi prisotnosti na mitohondrijskem genomu, kar bi lahko privedlo do hitrejših in bolj specifičnih zaznav patogenov neposredno iz zemeljskih vzorcev. Molekularni marker s podobnimi lastnostmi, ki temelji na ohranjenih PCR začetnih oligonukleotidih v mitohondrijski regiji za gen majhne rRNA podenote, je že bil predlagan za *V. dahliae* (Li in sod., 1994).

Za mutacijsko analizo jedrnih genov in filogenetsko analizo sevov smo izhajali iz variant s pripisanimi značilnostmi. Vhodne podatke smo pripravili na podlagi konkateniranih variant, pri čemer smo želeli zajeti čimveč variabilnosti v podatkih. Za ta namen smo uporabili sposobnost orodja SAMtools, da hkrati analizira podatke iz več vzorcev (v našem primeru vseh 6 vzorcev hkrati in ne vsakega posebej), kar zelo poveča sposobnost zaznave variant v regijah z nizko pokritostjo. To je možno zaradi korelacije med vzorci, kjer se lahko določi SNP, če se ta pojavi v več vzorcih, vendar je signal zaznave prešibek pri posameznem vzorcu. Od teh pridobljenih variant smo nato želeli pridobiti le visoko kakovostne, zato smo od zaznanih variant uporabili samo zamenjave, katerim se lahko enolično določi izvor v poravnavi variant vseh vzorcev. Od teh smo odfiltrirali variante, katerih genotipi so imeli enake verjetnosti ter variante, katerih vrednosti za kakovost genotipov so bile nizke. Pričakovali smo večje število variant od dobljenih, podobno kot se je izkazalo pri raziskavi 11-ih sevov sorodne fitopatogene glive *V. dahliae* (de Jonge in sod., 2012). V tej raziskavi je najsorodnejši sev sicer vseboval le 5.445 SNP (JR2) v primerjavi z referenčnim genomom, vendar se je na podlagi vseh 11-ih sevov določilo skupno 236.785 SNP pozicij. Pri naših sevih se je izkazalo da so si genomi precej bolj podobni, saj smo ugotovili prisotnost le 1.337 variant, od tega 1.080 SNP pozicij.

Iz slike naše filogenetske analize (Slika 7) je razvidno, da se skupini letalnih in blagih sevov ločita z visoko 100 % bootstrap podporo. Sev T2 je bil za grafični prikaz uporabljen kot izhodišče drevesa, vendar rezultati analize nakažejo, da je to drevo, pridobljeno po metodi največje verjetnosti, brez izhodišča (*angl. un-rooted tree*). Ker analiza le-tega ni mogla zagotovo določiti, potem tudi ne moremo potrditi ali ovreči domneve o nedavnem enotnem predniku, iz katerega so se razvili letalni sevi. Drug podatek, ki ga lahko dobimo iz tega drevesa, je izrazita ločnica med blagimi in letalnimi sevi ne glede na geografsko poreklo. Na podlagi tega sklepamo, da evolucija sevov ni potekala ločeno od ostalih na različnih geografskih področjih in da to vseeno nakazuje k skupnemu predniku, vendar le tega s to analizo oz. razpoložljivimi podatki nismo uspeli določiti.

Pri obravnavi genoma smo se odločili, da bomo genomske značilnosti (eksone, ponovitve in variante) pregledali ne samo preko pripisanih značilnosti, shranjenih v datotekah s pripisanimi značilnostmi, ampak tudi vizualno za obstoj morebitnih posebnih regij, ki bi vsebovale odstopanja v vsebnosti v primerjavi z ostalim genomom. Ena izmed motivacij je bila vizualna določitev morebitnih regij, kjer se pojavlja veliko variant v ozkem območju kodirajočih zaporedij. To vizualno analizo smo uresničili s prikazom gostotnih porazdelitev po genomu z drsnim oknom arbitrarno določene velikosti 10 kbp. Prva dilema, na katero smo naleteli ob tem, je bila normalizacija podatkov znotraj drsnega okna. Na voljo sta bili 2 možnosti: normalizacija glede na število genomskih značilnosti ali pa normalizacija glede na % pokritosti okna z njimi. Prednost prve je ta, da je za implementiranje precej preprosta, ima pa slabost, da daje večjo težo območjem, kjer je veliko število majhnih genomskih značilnosti in lahko včasih tudi izpusti mesta, kjer je teh malo, vendar so skupno glede na število baznih parov velike (npr. dolg zvezen niz ponovitvenih zaporedij). Ker nismo želeli izpustiti takšnih območij iz naše analize eksonov in ponovitvenih regij, smo se odločili implementirati pristop, ki upošteva % pokritosti okna. Ta pristop smo nato uporabili na vsakem posameznem kromosomu referenčnega seva *V. nonalfalfae* T2.

Slike gostotnih porazdelitev variant (Poglavje 4.2.2.) so zastavljene tako, da so trije letalni sevi na desni strani slike (1985, p15, T2) in trije blagi sevi na levi strani slike (1953, p55, rec91). Iz njih lahko razberemo, da imajo sevi znotraj patotipov (letalni, blagi) veliko skupnih variant in malo število odstopanj znotraj teh skupin. Opazno je tudi splošno nizko število variant v celotnem genomu, brez posebnih "vročih regij" obogatenih z mutacijami (največje št. mutacij v 10 kbp regiji beleži sev rec91 na kromosomu 4, položaju cca. 3.200.000 bp s 24 mutacijami). Slike gostotne porazdelitve ponovitvenih regij nam pokažejo relativno nizko stopnjo ponovitvenih zaporedij po kromosomih (v povprečju cca. 5-10 %) z nekaj precej visokimi odstopanji do 80 %, kar

je lahko posledica dejanskih ponovitev ali pa morebitnih napak pri združevanju sosesk v ogrodja (Hunt in sod., 2014). Nekateri kromosomi (npr. kromosom 4) imajo na obeh koncih viden majhen skok v gostoti porazdelitve ponovitev, kar nakazuje na prisotnost telomernih zaporedij in posledično na pravilno sestavljen kromosom (Faino in sod., 2015). Regije, kjer gostota porazdelitve ponovitvenih zaporedij pade na 0 %, so vezne regije z neinformativnim znakom N, ki predstavljajo manjkajoča področja, potrjena z optičnim kartiranjem. Pri gostotah porazdelitev eksonskih regij lahko vidimo precejšnjo razpršitev vrednosti, ki variira okoli povprečja 50 % in dosežene vrednosti ponekod segajo tudi do 100 %.

Pri postopku za analizo Ka/Ks koeficientov smo se srečali z nekaterimi dilemami, ki so vplivale na izvedbo analize. Največji problem te analize je bil, da je odvisna od kakovosti poravnave in lahko že manjše spremembe vplivajo na celoten preostanek poravnanega zaporedja od neke točke navzdol (Jordan in Goldman, 2011). Morebitnih protokolov za to analizo je bilo več, od katerih so se določeni distancirali od postopka poravnave in zahtevali, da se poravnava naredi drugje ter uvozi v postopek naknadno (Yang, 2007; Suzuki, 2011). Tisti, ki so vsebovali postopek poravnave, pa so zagovarjali različne pristope k temu problemu, kot npr. poravnavo na nivoju aminokislin (Reboiro-Jato in sod., 2012; Suyama in sod., 2006) ali pa poravnavo na nivoju kodonov (Zhang in sod., 2013; Ranwez in sod., 2011).

Poravnava na nivoju aminokislin poteka v treh korakih:
1) Prevod nukleotidnih zaporedij v proteinska zaporedja;
2) Poravnava proteinskih zaporedij;
3) Poravnava DNA zaporedij glede na proteinsko zaporedje.

Ta tro-stopenjski pristop je predvsem dovzeten za napake ob zamikih okvirja, katere dostikrat algoritem ne more enolično razrešiti. Če imamo vhodne kodirajoče podatke v obliki nukleotidnih zaporedij, potem moramo najprej ta prevesti v proteinska zaporedja, pri čemer nastopi problem predvidenega zaporedja proteina, ki ga pridobimo s prevajanjem iz nukleotidnega zaporedja in dejanskega proteinskega zaporedja. Precej podoben problem je tudi prevod proteinskega nazaj v nukleotidno zaporedje, kjer nastopi problem redundance genetskega koda.

Poravnava na nivoju kodonov predstavlja alternativo prejšnjemu pristopu, kjer se DNA zaporedja neposredno poravnajo z upoštevanjem kodonskih informacij. Implementacijo tega pristopa (Zhang in sod., 2013) smo uporabili v našem delu za poravnave in izračune Ka/Ks koeficientov jedrnih genov *V. nonalfalfae*. Ta temelji na uporabi programov BLAT (Kent, 2002) in bl2seq (Camacho in sod., 2009), katera poravnata

kodirajoča zaporedja iz referenčnega genoma na drug genom in ob tem odstranita nehomologna zaporedja kot so npr. pari delno podvojenih genov. S tem pristopom se upošteva morebitni obstoj večih mutacij premaknitve okvirja (*angl. frameshift*) in/ali zgodnjih stop kodonov ob poravnavi.

Pri naši analizi smo kot končni nabor pridobili 142 genov pod vplivom pozitivne selekcije. Ti izvirajo iz podatkov Ka/Ks analize sorodnih *Verticillium* vrst s programoma OrthoMCL in gKaKs. V obeh primerih so bili geni pod vplivom pozitivne selekcije pridobljeni na podlagi NG in YN metod ter so pri obeh imeli Ka/Ks vrednosti višje od 1. Čeprav je YN metoda strožja od NG, v povprečju pri naših rezultatih ni nakazovala nižjih Ka/Ks vrednosti v primerjavi z NG metodo.

Vrednosti prav tako ne nakazujejo visoke dovzetnosti genov za pozitivno selekcijo, pri povprečni Ka/Ks vrednosti 1,430 in standardnem odklonu 0,726 in tudi posebnih vzorcev porazdelitve teh genov po genomu nismo zaznali (Preglednica 15).

Za širšo sliko vloge genov, katerim smo določili da so pod pritiskom pozitivne selekcije, smo uporabili obogatitveno analizo GO pojmov. Ta nam je prikazala 26 skupin iz vseh treh glavnih vej genske ontologije (biološki proces, celična komponenta, molekularna funkcija), pri čemer je za vsako ocenila statistično značilno verjetnost obogatenosti ali pa osiromašenosti. Razberemo lahko, da geni pod vplivom pozitivne selekcije kažejo statistično značilno obogatenost v skupinah metabolnih procesov amino- in maščobnih kislin, fuzije veziklov z Golgijevim aparatom ter vnosom proteinov v mitohondrije pri skupini bioloških procesov. Prav tako kažejo obogatenost celičnih struktur, ki sodelujejo pri transportu snovi v in znotraj celice, kot npr, dineinski ter kinezinski kompleks, SNARE kompleks, membrana citoplazemskih veziklov (Preglednica 16). Te omenjene strukture so povezane s procesom fuzije veziklov z membranami in prenosom snovi pri rastlinskih patogenih (Schulze-Lefert, 2004; Leborgne-Castel, 2014; Kwon in sod., 2008) in glede na rezultate domnevamo, da so te strukture pri *V. nonalfalfae* nagnjene k pospešeni evoluciji za specializacijo vnosa efektorskih molekul v gostiteljsko rastlino. Med molekularnimi funkcijami opazimo tudi obogatenost proteinov signalnih procesov in procesa deiminacije oz. citrulinacije, vendar zaradi splošne vloge teh proteinov ne moremo sklepati, da so pod vplivom pozitivne selekcije samo zaradi njihove vloge pri povečani patogenosti *V. nonalfalfae.* Prisotne so tudi skupine genske ontologije, katerim je pripisana statistična značilna verjetnost osiromašenosti pojmov, kot npr. redoks procesi, razvoj micelija in vezave proteinov, vendar so to zelo splošno opredeljene skupine, ki se nahajajo visoko v drevesu genske ontologije in jih je težje obravnavati specifično v kontekstu patogenosti.

6 SKLEPI

V tem delu smo preučili določene genomske lastnosti fitopatogene glive *V. nonalfalfae* z bioinformatičnimi pristopi. Namen opravljenih raziskav je bil ugotoviti dejavnike, ki povzročajo povišano stopnjo patogenosti pri nekaterih *V. nonalfalfae* sevih na osnovi mitohondrijskega genoma in stopnje mutacij znotraj jedrnih genov.

Eden izmed rezultatov te raziskave je dokončan krožni 26.139 bp genom mitohondrija s podobnimi lastnostmi, kot so prisotne pri askomicetnih glivnih mitohondrijskih genomih, še posebej pri sorodni vrsti *V. dahliae*. Genom nosi zapise za 14 protein-kodirajočih genov, ki so vpleteni v procese oksidativne-fosforilacije ter sinteze ATP, poleg njih pa tudi malo in veliko rRNA podenoto, ribosomalni protein S3, ki se nahaja znotraj introna tipa-IA in 26 tRNA genov, kateri so skupna značilnost sordariomicetnih mitohondrijskih genomov. Dodatno smo v regiji, ki je bila že predhodno uporabljena v študiji za razlikovanje med izolati *V. dahliae* (EPPO Bulletin, 2007), odkrili še prisotnost potencialne dolge ne-kodirajoče RNA (*orf414*). To regijo smo našli samo pri *V. dahliae* in *V. nonalfalfae* in jo zaradi tega šteli za unikatno lastnost teh dveh vrst. Na podlagi proteinskih zaporedij s pripisanimi značilnostmi za rekonstrukcijo filogenetskega drevesa smo potrdili, da se naš mitohondrijski genom uvrsti v skupino gliv *Glomerellales* skupaj z *V. dahliae,* z zelo ohranjeno sintenijo. Uvrstitev v omenjeno skupino je podprta tudi na podlagi razvrščanja (*angl. clustering*) glede na podatke jedrnih genomov (Morgenstern in sod., 2012). Še ena zanimiva lastnost je dodatno zaporedjev *V. dahliae*, ki ni prisotno pri *V. nonalfalfae* in bi lahko bilo uporabljeno kot potencialni biomarker za razlikovanje med tema dvema vrstama.

Z analizami jedrnega genoma smo poskusili pojasniti evolucijski razvoj patogenih sevov *V. nonalfalfae*. Prvi korak k temu je bila določitev variant, pri čemer smo ugotovili, da se te pojavljajo v relativno nizkih nivojih (1.337 variant v celotnem *V. nonalfalfae* genomu); na podlagi njihovih gostotnih porazdelitev smo tudi ugotovili, da se na genomu ne pojavljajo vroče točke (*angl. hot-spots*) z visokimi koncentracijami variant ter da je gostotna porazdelitev eksonskih regij precej razpršena po genomu (niso opazne gruče povezanih kodirajočih enot - npr. povezane skupine genov, katere sestavljajo patogenske otočke). S pomočjo variant smo izvedli tudi filogenetsko analizo, ki je pokazala, da obstaja med sevi jasna ločnica pripadnosti skupinama letalnega in blagega patotipa. Ista analiza je prav tako nakazala, da na podlagi teh podatkov ni bilo mogoče določiti korenine (*angl. root*) filogenetskega drevesa, kar pomeni, da ni bilo možno enolično določiti izvorni sev *V. nonalfalfae,* iz katerega bi se razvili ostali.

V okviru analize evolucijskega pritiska pozitivne selekcije na jedrne gene *V. nonalfalfae* smo odkrili 142 genov, ki niso kazali posebnih porazdelitvenih vzorcev po genomu.

Obogatitvena analiza pojmov genske ontologije teh genov je pokazala statistično značilno obogatenost v skupinah, ki sodelujejo pri transportu snovi v celice in znotraj njih, iz česar sklepamo, da so te strukture pri *V. nonalfalfae* nagnjene k pospešeni evoluciji za specializacijo vnosa efektorskih molekul v gostiteljsko rastlino.

7 POVZETEK (SUMMARY)

7.1 POVZETEK

Tehnološki napredek na področju bioinformatike je omogočil nove pristope analiz genomskih podatkov z uporabo NGS tehnologij, vzporedno z njim pa je splošen napredek na področju računalništva, tako strojne kot tudi programske opreme, ustvaril pogoje sinergije, ki so spodbudili razvoj obeh področij ter privedli do tesnejših nivojev sodelovanja, kateri so opazni v čedalje bolj celostnih in kompleksnih obravnavah bioloških podatkov. V naši raziskavi smo uporabili nekatere izmed takšnih metod celostnega obravnavanja biološkega sistema za analizo genomskih podatkov *V. nonalfalfae* in pri tem pripravili svoje lastno programsko ogrodje za izvedbo izbranih analiz.

Namen analize fitopatogene glive *V. nonalfalfae* je bil ugotovitev dejavnikov, ki povzročajo povišano stopnjo patogenosti sevov, katere smo pridobili iz treh različnih geografskih lokacij in jih preučili z metodami, ki so vsebovale analize mitohondrijskega genoma ter analize mutacij jedrnih genov. Z analizo mitohondrijskih zaporedij v naboru NGS podatkov smo sestavili končni 26.139 bp velik krožni mitohondrijski genom, ki je po svojih lastnostih zelo podoben mitohondrijskemu genomu *V. dahliae*. Na njem se nahaja 14 protein-kodirajočih genov, mala in velika rRNA podenota, ribosomalni protein S3 znotraj introna tipa-IA in 26 tRNA genov, kar so pogoste karakteristike sordariomicetnih mitohondrijskih genomov. Poleg teh smo odkrili tudi 2 dodatni lastnosti: potencialno dolgo ne-kodirajočo RNA (*orf414*), ki smo jo zaznali samo pri *V. dahliae* in *V. nonalfalfae* ter dodatno zaporedje v *V. dahliae*, ki ni prisotno pri *V. nonalfalfae* in bi lahko bilo uporabljeno kot potencialni biomarker za razlikovanje med tema dvema vrstama. Ob analizah jedrnega genoma smo obravnavali gostotne porazdelitve variant, eksonskih ter ponovitvenih regij po genomu, pri čemer se je izkazalo, da ima naš *V. nonalfalfae* genom relativno nizko raven variant - 1.337 variant, ki se ne pojavljajo v večjih gručah, da sta gostotni porazdelitvi eksonskih ter ponovitvenih regij precej razpršeni po genomu in da porazdelitev variant nima visokega vpliva na kodirajoča zaporedja. Na teh variantah temelječa filogenetska analiza je pokazala, da obstaja med letalnim in blagim patotipom jasna ločnica in da na podlagi teh podatkov ni bilo mogoče določiti korenine filogenetskega drevesa in s tem izvornega seva *V. nonalfalfae,* iz katerega bi se domnevno razvili ostali. Poleg te filogenetske analize med sevi smo izvedli še filogenetsko analizo med vrstami na podlagi mitohondrijskega genoma, s katero smo ugotovili, da se naš mitohondrijski genom uvršča v skupino gliv *Glomerellales* skupaj z *V. dahliae* z zelo ohranjeno sintenijo, kar se je izkazalo za skladno tudi z rezultati razvrščanja glivnih vrst glede na podatke jedrnih genomov. Z evolucijsko analizo jedrnih genov smo našli 142 genov

podvrženih pozitivni selekciji in preko obogatitvene analize pojmov genske ontologije teh genov ugotovili, da so obogatene tiste skupine, ki so vpletene v transport snovi v celice in znotraj celic, iz česar sklepamo da se je genom letalnega patotipa *V. nonalfalfae* preko pozitivnega selekcijskega pritiska adaptiral k optimizaciji transporta efektorskih molekul.

7.2 SUMMARY

Technological advances in the field of bioinformatics have led to new approaches towards genome data analysis, through the use of NGS technologies. Parallel progress in the field of computation technology, both in software and hardware, has led to newly established synergies, which stimulated further development of both fields and resulted in tighter collaboration efforts. These collaborations in turn have brought forth more systemic and complex studies of biological data. In this work, we have used some of the multi-level encompassing methods to analyze genomic data of *V. nonalfalfae* and alongside developed our own software framework for conducting customized analyses.

The goal of the *V. nonalfalfae* pathogenic fungi analysis was to determine the factors, which give rise to a heightened level of strain pathogenicity for the strains we gathered from three distinct geographical areas. This was accomplished with methods that encompassed mitochondrial genome analysis and mutation analysis of nuclear genes. By analyzing mitochondrial sequences in the NGS datasets, we have assembled a final 26.139 bp mitochondrial genome, that bears features reminiscent of the *V. dahliae* mitochondrial genome. It harbors 14 protein-coding genes, a small and large rRNA subunit, a ribosomal protein S3 inside an IA-type intron and 26 tRNA genes, which are common characteristics of sordariomycete mitochondrial genomes. We further discovered 2 additional features: a potential long non-coding RNA (*orf414*), which we traced only by *V. dahliae* and *V. nonalfalfae* and an additional sequence in *V. dahliae,* that is not present in *V. nonalfalfae,* which could be potentially used as a biomarker to distinguish between these two species. During analyses of the nuclear genome, we took into consideration distribution densities of variants, exonic and repetitive regions across the genome, which have shown that our *V. nonalfalfae* genome has a relatively low level of variants - 1.337 variants, which do not occur in large clusters, that the exon and repeat densities are uniformly spread along the genome and that the distribution of variants does not have a major impact on the coding sequences. A phylogenetic analysis based on these variants has shown, that there is a clear boundary between the lethal and mild pathotype and that based on this data, it is not possible to determine the phylogenetic tree root and consequentially the source strain of *V. nonalfalfae,* from which the other strains would allegedly evolve. Along-side this intra-species analysis

we have also conducted a inter-species analysis, by which we discovered that our mitochondrial genome is placed in the *Glomerellales* fungal group, together with *V. dahliae* bearing a highly conserved synteny. These results are also in line with the nuclear genome-based clustering of fungal species mentioned earlier in this work. With an evolutionary analysis of nuclear genes, we have found 142 genes subjected to positive selection. A gene ontology enrichment analysis of these genes has showed that the enriched groups are the ones who are connected to matter transport into and within the cells. This led us to a conclusion, that the *V. nonalfalfae* genome of the lethal pathotype has optimised its effector molecule transport through positive selection adaptations.

8 VIRI

Al-Reedy R. M., Malireddy R., Dillman C. B., Kennell J. C. 2012. Comparative analysis of *Fusarium* mitochondrial genomes reveals a highly variable region that encodes an exceptionally large open reading frame. Fungal Genetics and Biology, 49, 1: 2–14

Anders S., Huber W. 2010. Differential expression analysis for sequence count data. Genome Biology, 11: R106

Bakker F. T., Lei D., Yu J., Mohammadin S., Wei Z., van de Kerke S., Gravendeel B., Niewenhuis M., Staats M., Alquezar-planas D. E., Holmer R. 2016. Herbarium genomics: plastome sequencing assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. Biological Journal of the Linnean Society, 117, 1: 33-43

Bendtsen J. D., Nielsen H., Heijne Gv., Brunak S. 2004. Improved prediction of signal peptides: SignalP. 3.0. Journal of Molecular Biology, 340, 4: 783-795

Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Wheeler D. L. 2005. GenBank. Nucleic Acids Research, 33, supplement 1: D34-D38, doi: 10.1093/nar/gki063: 5 str.

Besemer J., Lomsadze A., Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Research, 29, 12, doi: 10.1093/nar/29.12.2607: 2607-2618

Bofkin L., Goldman N. 2007. Variation in evolutionary processes at different codon positions. Molecular Biology and Evolution, 24, 2: 513-521

Borodovsky M., McIninch J. 1993. GeneMark: parallel gene recognition for both DNA strands. Computers & Chemistry, 17, 19: 123-133

Bright L. A., Burgess S. C., Chowdhary B., Swiderski C. E., McCarthy F. M. 2009. Structural and functional-annotation of an equine whole genome oligoarray. BMC Bioinformatics, 10, doi: 10.1186/1471-2105-10-S11-S8: 8 str.

Burger G., Forget L., Zhu Y., Gray M. W., Lang B. F. 2003. Unique mitochondrial genome architecture in unicellular elatives of animals. Proceedings of the National Academy of Sciences, 100, 3: 892-897

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T. L. 2009. BLAST+: architecture and applications. BMC Bioinformatics, 10: 421, doi: 10.1186/1471-2105-10-421: 9 str.

Cantarel B. L., Korf I., Robb S. M. C., Parra G., Ross E., Moore B. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research, 18, 1: 188–196

Capella-Gutiérrez S., Silla-Martínez J. M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25, 15: 1972-1973

Cardoso M. A., Tambor J. H., Nobrega F. G. 2007. The mitochondrial genome from the thermal dimorphic fungus *Paracoccidioides brasiliensis*. Yeast, 24, 7: 607–616

Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S. J., Lu X., Ruden D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Landes Bioscience, 6, 2: 80-92

Clark S. C., Egan R., Frazier P. I., Wang Z. 2013. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics, 29, 4: 435-443

Conesa A., Götz S., Garcia-Gomez J. M., Terol J., Talon M., Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21, 18: 3674–3676

Cristianini N., Hahn M. W. 2006. Introduction to computational genomics: a case studies approach. New York, Cambridge University Press: 202 str.

Danecek P., Auton A., Abecaseis G., Albers C. A., Banks E., DePristo M. A., Handsaker R. E., Lunter G., Marth G. T., Sherry S. T., McVean G., Durbin R. 2011. The variant call format and VCFtools. Bioinformatics, 27, 15: 2156-2158

Darty K., Denise A., Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics, 25, 15: 1974-1975

Dong S., Raffaele S., Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. Current Opinion in Genetics and Development, 35:57-65, doi: 10.1016/j.gde.2015.09.001: 9 str.

Duressa D., Anchieta A., Chen D., Klimes A., Garcia-Pedrajas M. D., Dobinson K. F., Klosterman S. J. 2013. RNA-seq analyses of gene expression in the microsclerotia of Verticillium dahliae. BMC Genomics, 14: 607, doi: 10.1186/1471-2164-14-607: 18 str.

Eddy S. R. 2011. Accelerated profile HMM searches. PLoS Computational Biology, 7, 10, doi: 10.1371/journal.pcbi.1002195: 16 str.

Edgar R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. Nucleic Acids Research, 32, 5: 1792–1797

Ekblom R., Wolf J. B. W. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications, 7, 9: 1026-1042

El-Metwally S., Hamza T., Zakaria M., Helmy M. 2013. Next-generation sequence assembly: Four stages of data processing and computational challenges. PLoS Computational Biology, 9, 12: e1003345, doi:10.1371/journal.pcbi.1003345: 19 str.

Faino L., Seidl M. F., Datema E., van den Berg G. C., Janssen A., Wittenberg A. H., Thomma B. P. 2015. Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. mBio, 6, 4: e00936-15, doi:10.1128/mBio.00936-15: 11 str.

Gautheret D., Lambert A. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. Journal of Molecular Biology, 313, 5: 1003-1011

Gent D. H., Woods J. L., Putnam M. L. 2012. New outbreaks of Verticillium wilt on hop in Oregon caused by *Verticillium albo-atrum*. Plant Health Progress, 13, doi:10.1094/PHP-2012-0521-01-RS: 10 str.

Giegerich R. 2011. Introduction to stochastic context free grammars: 23 str. http://www.techfak.unibielefeld.de/ags/pi/lehre/RNA_StrukturSS11/IntroStochGram.pdf (11. dec. 2016)

Goto H., Dickins B., Afgan E., Paul I. M., Taylor J., Makova K. D., Nekrutenko A. 2011. Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biology 2011, 12, 6: R59

Gurevich A., Saveliev V., Vyahhi N., Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29, 8: 1072-1075

Haas B. J., Volfovsky N., Town C. D., Troukhan M., Alexandrov N., Feldmann K. A., Flavell R. B., White O., Salzberg S. L. 2002. Full-length messenger RNA sequences greatly improve genome annotation. Genome Biology, 3, 6: 1-12

Haas B. J., Zeng Q., Pearson M. D., Cuomo C. A., Wortman J. R. 2011. Approaches to fungal genome annotation. Mycology, 2, 3: 118-141

Haas B. 2015. TransDecoder. http://transdecoder.github.io/ (11. dec. 2016)

Haibao T., Pedersen B., Ramirez F., Naldi A., Flick P., Yunes J., Sato K., Mungall C., Stupp G., Klopfenstein D. V., DeTomaso D., Botvinnik O. 2015. GOATOOLS: Tools for gene ontology. Zenodo. http://dx.doi.org/10.5281/zenodo.31628 (15.5.2016)

Holt C., Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics, 12: 491, doi:10.1186/1471-2105-12-491: 14 str.

Hunt M., Newbold C., Berriman M., Otto T. D. 2014. A comprehensive evaluation of assembly scaffolding tools. Genome Biology 2014, 15, 3: R42, doi**:** 10.1186/gb-2014-15-3-r42: 15 str.

Huynen M., Snel B., Lathe W., Bork P. 2000. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. Genome Research: 10, 8: 1204-1210

Inderbitzin P., Bostock R. M., Davis R. M., Usami T., Platt H. W., Subbarao K. V. 2011. Phylogenetics and taxonomy of the fungal vascular wilt pathogen Verticillium, with the descriptions of five new species. PloS ONE, 6, doi: 10.1371/journal.pone.0028341: 22 str.

Javornik B. 2012. Genom *V. albo-atrum*. (osebni vir, 10. 11. 2012)

Jelen V., de Jonge R., Van de Peer Y., Javornik B., Jakše J. 2016. Complete mitochondrial genome of the Verticillium-wilt causing plant pathogen *Verticillium nonalfalfae*. PLoS ONE 11, 2: e0148525, doi:10.1371/journal.pone.0148525: 18 str.

Jordan M., Goldman N. 2011. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Molecular Biology and Evolution, 29, 4: 1125-1139

Jung P. P., Friedrich A., Reisser C., Hou J., Schacherer J. 2012. Mitochondrial genome evolution in a single protoploid yeast species. G3 (Bethesda), 2, 9, doi: 10.1534/g3.112.003152: 1103-1111

Kent W. J. 2002. BLAT - The BLAST-Like alignment tool. Genome Research, 12, 4: 656-664

Keyworth W. G. 1942. Verticillium wilt of the hop (*Humulus lupulus*). Annals of Applied Biology, 29, 4: 346-357

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., Salzberg S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology, 14: R36, doi:10.1186/gb-2013-14-4-r36: 13 str.

Klosterman S. J., Subbarao K. V., Kang S., Veronese P., Gold S. E., Thomma Bp H. J., Chen Z., Henrissat B., Lee Y. H., Park J., Garcia-Pedrajas M. D., Barbara D. J., Anchieta A., de Jonge R., Santhanam P., Maruthachalam K., Atallah Z., Amyotte S. G., Paz Z., Inderbitzin P., Hayes R. J., Heiman D. I., Young S., Zeng Q., Engels R., Galagan J., Cuomo C. A., Dobinson K. F., Ma L. 2011. Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. PLoS Pathogens 7, 7: e1002137, doi:10.1371/journal.ppat.1002137: 19 str.

Knudsen B., Hein J. 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics, 15, 6: 446-454

Kosa P., Valach M., Tomaska L., Wolfe K. H., Nosek J. 2006. Complete DNA sequences of the mitochondrial genomes of the pathogenic yeasts *Candida orthopsilosis* and *Candida metapsilosis*: insight into the evolution of linear DNA genomes from mitochondrial telomere mutants. Nucleic Acids Research, 34, 8: 2472-2481

Kouvelis V. N., Ghikas D. V., Typas M. A. 2004. The analysis of the complete mitochondrial genome of *Lecanicillium muscarium* (synonym *Verticillium lecanii*) suggests a minimum common gene organization in mtDNAs of *Sordariomycetes*: phylogenetic implications. Fungal Genetics and Biology, 41, 10: 930–940

Krzywinski M., Schein J. E., Birol I., Connors J., Gascoyne R., Horsman D., Jones S. J., Marra M. A. 2009. Circos: an information aesthetic for comparative genomics. Genome Research, 19, 9: 1639–1645

Kurtz S., Phillippy A., Delcher A. L., Smoot M., Shumway M., Antonescu C., Salzberg S. L. 2004. Versatile and open software for comparing large genomes. Genome Biology 5, 2: R12, doi: 10.1186/gb-2004-5-2-r12: 13 str.

Kwon C., Bednarek P., Schulze-Lefert P. 2008. Secretory pathways in plant immune responses. Plant Physiology, 147, 4, doi: 10.1104/pp.108.121566: 1575-1583

Langmead B., Trapnell C., Pop M., Salzberg S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, 3: R25, doi: 10.1186/gb-2009-10-3-r25: 10 str.

Larran P., Calvo B., Santana R., Bielza C., Galdiano J., Inza I., Lozano J. A., Armananzas R., Santafe G., Perez A., Robles V. 2005. Machine learning in bioinformatics. Briefings in Bioinformatics, 7, 1: 86-112

Laslett D., Canbäck B. 2008. ARWEN, a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics, 24, 2: 172-175

Lee E., Helt G. A., Reese J. T., Munoz-Torres M. C., Childers C. P., Buels R. M., Stein L., Holmes I. H., Elsik C. G., Lewis S. E. 2013. Web Apollo: a web-based genomic annotation editing platform. Genome Biology, 14: R93 doi:10.1186/gb-2013-14-8-r93: 13 str.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The sequence alignment/map format and SAMTools. Bioinformatics, 25, 16, doi: 10.1093/bioinformatics/btp352: 2078-2079

Li K. N., Rouse D. I., German T. L. 1994. PCR primers that allow intergenic differentiation of ascomycetes and their application to *Verticillium* spp. Applied Environmental Microbiology, 60, 12: 4324-31

Li L., Stoeckert C. J., Roos D. S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Research, 13, 9, doi: 10.1101/gr.1224503: 2178-2189

Lowe T. M., Eddy S. R. 1997. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25, 5: 955-964

MacLean D., Jones J. D. G., Studholme D. J. 2009. Application of 'next-generation' sequencing technologies to microbial genetics. Nature Reviews Microbiology, 7, 4: 287-296

Madoui M. A., Dossat C., Agata L., van Oeveren J., van der Vossen E., Aury J. M. 2016. MaGuS: a tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling™ Data. BMC Bioinformatics, 2016, 17: 115, doi:10.1186/s12859-016-0969-x: 9 str.

Manning T., Sleator R. D., Walsh P. 2013. Naturally selecting solutions: The use of genetic algorithms in Bioinformatics. Bioengineered, 4, 5: 266-278

Mardanov A. V., Beletsky A. V., Kadnikov V. V., Ignatov A. N., Ravin N. V. 2014. The 203 kbp mitochondrial genome of the phytopathogenic fungus *Sclerotinia borealis* reveals multiple invasions of introns and genomic duplications. PLoS ONE, 9, 9: e107536, doi:10.1371/journal.pone.0107536: 11 str.

Martin F. N. 2010. Mitochondrial haplotype analysis as a tool for differentiating isolates of *Verticillium dahliae*. Mycology, 100, 11, doi:10.1094/PHYTO-12-09-0352: 1231- 1239

Reboiro-Jato D., Reboiro-Jato M., Fdez-Riverola F., Vieira C. P., Fonseca N. A., Vieira J. 2012. ADOPS - Automatic detection of positively selected sites. Journal of Integrative Bioinformatics, 9, 3: 200, doi:10.2390/biecoll-jib-2012-200: 15 str.

Miller J. R., Koren S., Sutton G.. 2010. Assembly algorithms for next-generation sequencing data. Genomics, 95, 6: 315-327

Morgenstern I., Powlowski J., Ishmael N., Darmond C., Marqueteau S., Moisan M. C., Quenneville G., Tsang A. 2012. A molecular phylogeny of thermophilic fungi. Fungal Biology, 116, 4, doi: 10.1016/j.funbio.2012.01.010: 489-502

Pantou M. P., Typas M. A. 2005. Electrophoretic karyotype and gene mapping of the vascular wilt fungus Verticillium dahliae. FEMS Microbiology Letters, 245, 2: 213-220

Pantou M. P., Kouvelis V. N., Typas M. A. 2006. The complete mitochondrial genome of the vascular wilt fungus *Verticillium dahliae*: a novel gene order for *Verticillium* and a diagnostic tool for species identification. Current Genetics, 50, 2: 125–136

Petty N. K. 2010. Genome annotation: man versus machine. Nature Reviews Microbiology 8, 11: 762

Pevzner P. A., Tang H., Waterman M. S. 2001. An Eulerian path approach to DNA fragment assembly. Proceedings on the National Academy of Science USA, 98, 17: 9748–9753

Radišek S., Jakše J., Simončič A., Javornik B. 2003. Characterization of Verticillium albo-atrum field isolates using pathogenicity data and AFLP analysis. Plant Disease, 87, 6: 633-638

Ranwez V., Harispe S., Delsuc F., Douzery EJP. 2011 MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. PLoS ONE 6, 9: e22594, doi:10.1371/journal.pone.0022594: 10 str.

Rice P., Longden I., Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, 16, 6: 276-277

Roberts A., Pimentel H., Trapnell C., Pachter L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics, 27, 17: 2325-2329

Schadt E. 2010. A window into third-generation sequencing. Human Molecular Genetics, 19, 2, doi:10.1093/hmg/ddq416: 227-240

Schatz M. C., Delcher A. L., Salzberg S. L. 2010. Assembly of large genomes using second-generation sequencing. Genome Research, 20, 9: 1165-1173

Schulze-Lefert P. 2004. Knocking on the heaven's wall: pathogenesis of and resistance to biotrophic fungi at the cell wall. Current Opinion in Plant biology, 7, 4: 377-383

Seefelder S., Seigner E., Niedermeier E., Radišek S., Javornik B. 2009. Genotyping of Verticillium pathotypes in the Hallertau: basic findings to assess the risk of Verticillium infections. V: Proceedings of the Scientific Commision of the CICH - IHB - IHGC International Hop Growers' Convention, León, Spain, 21-25 June 2009. Seigner E. (ur.) Wolnzach: Scientific Comission, I.H.G.C: 67-69

Sethuraman J., Majer A., Iranpour M., Hausner G. 2009. Molecular evolution of the mtDNA encoded rps3 gene among filamentous ascomycetes fungi with an emphasis on the Ophiostomatoid fungi. Journal of Molecular Evolution, 69, 4, doi: 10.1007/s00239-009-9291-9: 372-385

Slater G. S. C., Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 6:31, doi:10.1186/1471-2105-6-31: 11 str.

Soanes D. M., Richards T. A., Talbot N. J. 2007. Genomes: what can we learn about plant insights from sequencing fungal and oomycete disease and the evolution of pathogenicity? Plant Cell, 19, 11: 3318-3326

Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics, 30, 9: 1312-1313

Stone C. L., Buitrago M. L., Boore J. L., Frederick R. D. 2010. Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens *Phakospora pachyrhizi* and *P. meibomiae*. Mycologia, 102, 4: 887–897

Suyama M., Torrents D., Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research, 1, 34: 609-612

Suzuki Y. 2011. Overestimation of nonsynonymous/synonymous rate ratio by reverse-translation of aligned amino acid sequences. Genes and Genetic Systems, 86, 2: 123–129

Torriani S. F. F., Penselin D., Knogge W., Felder M., Taudien S., Platzer M., McDonalnd B. A., Brunner P. C. 2014. Comparative analysis of mitochondrial genomes from closely related *Rhynchosporium* species reveals extensive intron invasion. Fungal Genetics and Biology, 62, doi:10.1016/j.fgb.2013.11.001: 9 str.

Tritt A., Eisen J. A., Facciotti M. T., Darling A. E. 2012. An integrated pipeline for de novo assembly of microbial genomes. PLoS ONE 7, 9: e42304, doi:10.1371/journal.pone.0042304: 9 str.

Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B. C., Remm M., Rozen S. G. 2012. Primer3 - new capabilities and interfaces. Nucleic Acids Research, 40, 15: 115

van de Sande W. W. J. 2012. Phylogenetic analysis of the complete mitochondrial genome of *Madurella mycetomatis* confirms its taxonomic position within the order *Sordariales*. PLoS ONE 7, 6: e38654, doi:10.1371/journal.pone.0038654: 10 str.

van Rossum G. 2001. Python programming language. http://www.python.org/ (11. dec. 2016)

Verticillium nonalfalfae and *V. dahliae* on hop. 2007. EPPO Bulletin, 37, 2, 528–535

Vitti J. J., Grossman S. R., Sabeti P. C. 2013. Detecting natural selection in Genomic data. Annual Review of Genetics, 47: 97-120

Yang P., Yang Y. H., Zhou B. B. 2010. A review of ensemble methods in bioinformatics. Current Bioinformatics, 5, 4: 296-308

Yang Z., Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution, 17, 1: 32-43

Yang Z. 2007. PAML 4: A program package for phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution, 24, 8: 1586-1591

Yang Z. R. 2004. Biological applications of support vector machines. Briefings in Bioinformatics, 5, 4: 328-338

Zerbino D. R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Current Protocols in Bioinformatics, 11: 11.5, doi: 10.1002/0471250953.bi1105s31: 13 str.

Zhang C., Wang J., Long M., Fan C. 2013. gKaKs: the pipeline for genome-level Ka/Ks calculation. Bioinformatics, 29, 5: 645-646

Zhang Z., Zhao X. Q., Wang J., Wong G. K. S., Yu J. 2006. KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics and Bioinformatics, 4, 4: 259-263

Zhao X. Q., Aizawa T., Schneider J., Wang C., Shen R. F., Sunairi M. 2013. Complete mitochondrial genome of the aluminium-tolerant fungus *Rhodotorula taiwanensis* RS1 and comparative analysis of *Basidiomycota* mitochondrial genomes. Microbiology Open, 2, 2: 308–317

ZAHVALA

Najprej bi se rad zahvalil svojemu mentorju: izr. prof. dr. Jerneju Jakšetu, ki mi je nudil veliko podporo in pomoč pri tem doktorskem delu ter prof. dr. Branki Javornik, ki mi je omogočila delo na tem projektu.

Za velik del znanja in dobrih praks programiranja bi se rad zahvalil: izr. prof. Janezu Demšarju, doc. dr. Tomažu Dobravcu ter asistentu Marku Toplaku, vsem iz FRI (Fakultete za računalništvo in informatiko, Univerze v Ljubljani).

I would also like to thank the BEG (Bioinformatics and Evolutionary Genomics) group of the PSB (Plant Systems Biology) at VIB (Flanders Institute for biotechnology) in Gent, Belgium, especially dr. Ronnie de Jonge, who mentored me during my research visit there.

Največja zahvala pa gre moji družini in vsem prijateljem, ki so me spodbujali in pri mojem delu podpirali.

PRILOGA A

Preverjeni *Verticillium spp.* izolati za dolžinski polimorfizem

Pomnožitev amplikona je označena v zadnjih dveh stolpcih s številko 1. Ena pomeni pomnožitev določene dolžine fragmenta, 0 pomeni brez pomnožitve. Če je v obeh stoplcih prisotna številko 0, potem do pomnožitve ni prišlo.

Evaluated *Verticillium spp.* isolates for the length polymorphism

Amplicon amplification is determined in the final 2 columns with the number 1. A one means the amplification of a fragment of certain length, a zero means no amplification. If both columns harbor 0, then the amplification did not take place.

			Velikost amplikona	
Št.	**Ime izolata**	**Vrsta**	**1400 bp**	**400 bp**
1	CBS 102.464	*V. alboatrum*	0	0
2	CBS 682.88	*V. alboatrum*	0	0
3	110	*V. alboatrum*	0	0
4	PD693	*V. alboatrum*	0	0
5	166	*V. alboatrum*	0	0
6	112	*V. alboatrum*	0	0
7	Luc	*V. alfalfae*	0	1
8	41	*V. alfalfae*	0	1
9	CBS 392.91	*V. alfalfae*	0	1
10	107	*V. alfalfae*	0	1
11	Kanada 11	*V. alfalfae*	0	1
12	CIG3	*V. dahliae*	1	0
13	JKG 2	*V. dahliae*	1	0
14	JKG1	*V. dahliae*	1	0
15	DJK	*V. dahliae*	1	0
16	MAI	*V. dahliae*	1	0
17	Mint	*V. dahliae*	1	0
18	GAJ09	*V. dahliae*	1	0
19	PDRENU/MAR	*V. dahliae*	1	0
20	CasD	*V. dahliae*	1	0
21	KresD	*V. dahliae*	1	0
22	MoD	*V. dahliae*	1	0
23	Oset	*V. dahliae*	1	0
24	12042	*V. dahliae*	0	0
25	PAP	*V. dahliae*	1	0
26	Pap99	*V. dahliae*	1	0
27	Pap2008	*V. dahliae*	1	0
28	14	*V. dahliae*	1	0
29	141	*V. dahliae*	0	0

se nadaljuje

nadaljevanje priloge A

Št.	Ime izolata	Vrsta	1400 bp	400 bp
30	3V	*V. dahliae*	1	0
31	802-1	*V. dahliae*	1	0
32	V 138 I	*V. dahliae*	1	0
33	V 176 I	*V. dahliae*	1	0
34	PD335	*V. dahliae*	1	0
35	PD584	*V. dahliae*	1	0
36	A III 25	*V. dahliae*	0	0
37	PD337	*V. dahliae*	0	0
38	JKG 20	*V. isaacii*	0	0
39	115	*V. isaacii*	0	0
40	EX5 F8	*V. isaacii*	0	0
41	CBS110218	*V. longisporum*	0	1
42	PD330	*V. longisporum*	0	1
43	P10	*V. nonalfalfae*	0	1
44	P114/1	*V. nonalfalfae*	0	1
45	P34/1	*V. nonalfalfae*	0	1
46	P15	*V. nonalfalfae*	0	1
47	P83	*V. nonalfalfae*	0	1
48	Jun-99	*V. nonalfalfae*	0	1
49	14/93	*V. nonalfalfae*	0	1
50	15/98	*V. nonalfalfae*	0	1
51	T2	*V. nonalfalfae*	0	1
52	TABOR 6	*V. nonalfalfae*	0	1
53	Ciz/DED	*V. nonalfalfae*	0	1
54	BIZ	*V. nonalfalfae*	0	1
55	MO 3	*V. nonalfalfae*	0	1
56	OCer	*V. nonalfalfae*	0	0
57	zup	*V. nonalfalfae*	0	1
58	Rec91	*V. nonalfalfae*	0	1
59	KRES 98	*V. nonalfalfae*	0	1
60	1985a	*V. nonalfalfae*	0	1
61	11055	*V. nonalfalfae*	0	1
62	11047	*V. nonalfalfae*	0	1
63	11100	*V. nonalfalfae*	0	1
64	1974	*V. nonalfalfae*	0	1
65	298102	*V. nonalfalfae*	0	1
66	1953	*V. nonalfalfae*	0	1
67	298092	*V. nonalfalfae*	0	0
68	298095	*V. nonalfalfae*	0	1

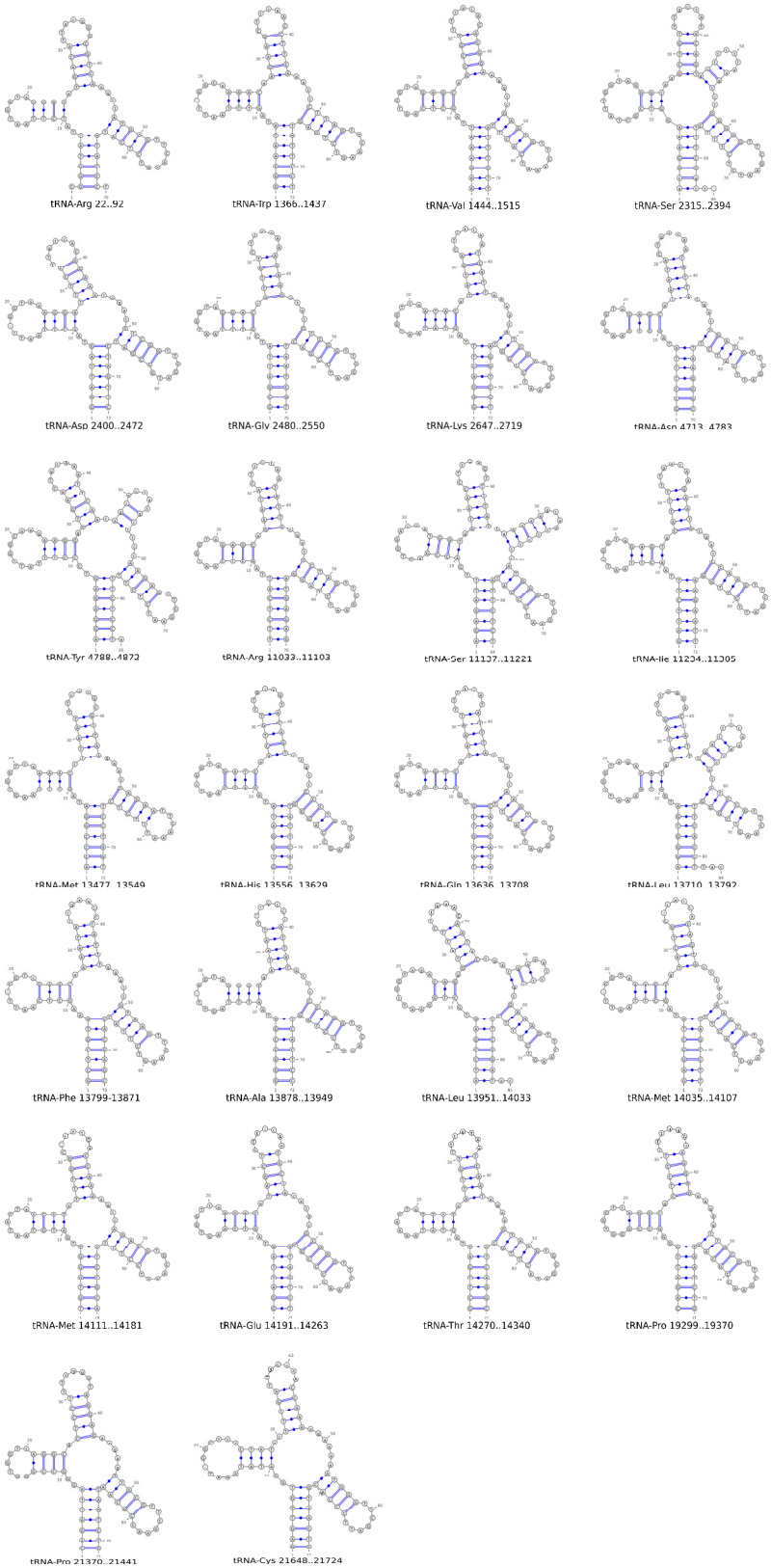
se nadaljuje

nadaljevanje priloge A

Št.	Ime izolata	Vrsta	1400 bp	400 bp
69	Sol	*V. nonalfalfae*	0	1
70	Surf	*V. nonalfalfae*	0	1
71	11081	*V. nonalfalfae*	0	1
72	11066	*V. nonalfalfae*	0	0
73	CBS 321.91	*V. nonalfalfae*	0	1
74	AR01/067	*V. nonalfalfae*	0	1
75	AR0/140	*V. nonalfalfae*	0	1
76	AR01/JS1	*V. nonalfalfae*	0	1
77	PD 83/53a	*V. nonalfalfae*	0	1
78	PD 2000/4186a	*V. nonalfalfae*	0	1
79	314193	*V. nonalfalfae*	0	1
80	SN 10	*V. nonalfalfae*	0	0
81	P84/2	*V. nonalfalfae*	0	1
82	KV11 URŠIČ	*V. nonalfalfae*	0	1
83	VranBis09	*V. nonalfalfae*	0	1
84	Sent4	*V. nonalfalfae*	0	1
85	11097	*V. nonalfalfae*	0	1
86	298100	*V. nonalfalfae*	0	0
87	kum	*V. nonalfalfae*	0	1
88	11077	*V. nonalfalfae*	0	1
89	CBS 454.51	*V. nonalfalfae*	0	1
90	340646	*V. nonalfalfae*	0	1
91	PETROL	*V. nonalfalfae*	0	1
92	PAPmb	*V. nigrescens*	0	0
93	CBS 123.176	*V. nigrescens*	0	0
94	CBS 456.51	*V. nubilum*	1	0
95	CBS 227.84	*V. tricorpus*	0	0
96	EX5 F7	*V. tricorpus*	0	0

PRILOGA B

Sekundarne strukture tRNA molekul v mitohondrijskem genomu
Secondary structures of tRNA molecules in the mitochondrial genome



PRILOGA C

Preglednica uporabe kodonov
Codon usage table

Kodon	Aminokislina	Delež	Število
GCA	A	0,224	46
GCC	A	0,112	23
GCG	A	0,034	7
GCU	A	0,629	129
UGC	C	0,34	54
UGU	C	0,66	105
GAC	D	0,318	41
GAU	D	0,682	88
GAA	E	0,737	98
GAG	E	0,263	35
UUC	F	0,27	148
UUU	F	0,73	401
GGA	G	0,411	104
GGC	G	0,055	14
GGG	G	0,111	28
GGU	G	0,423	107
CAC	H	0,301	37
CAU	H	0,699	86
AUA	I	0,39	311
AUC	I	0,129	103
AUU	I	0,481	384
AAA	K	0,685	231
AAG	K	0,315	106
CUA	L	0,083	66
CUC	L	0,033	26
CUG	L	0,052	41
CUU	L	0,163	129
UUA	L	0,585	463
UUG	L	0,083	66
AUG	M	1	124
AAC	N	0,285	112
AAU	N	0,715	281

se nadaljuje

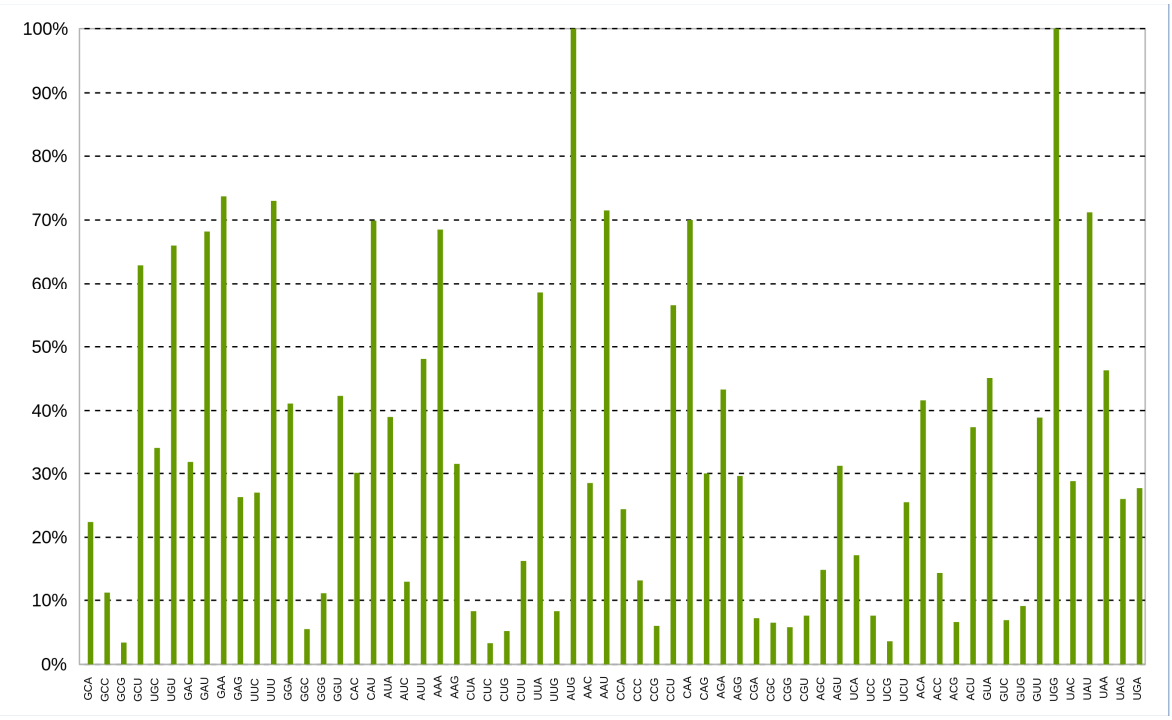
nadaljevanje priloge C

Kodon	Aminokislina	Delež	Število
CCA	P	0,244	41
CCC	P	0,131	22
CCG	P	0,06	10
CCU	P	0,565	95
CAA	Q	0,7	91
CAG	Q	0,3	39
AGA	R	0,433	120
AGG	R	0,296	82
CGA	R	0,072	20
CGC	R	0,065	18
CGG	R	0,058	16
CGU	R	0,076	21
AGC	S	0,149	79
AGU	S	0,312	165
UCA	S	0,172	91
UCC	S	0,076	40
UCG	S	0,036	19
UCU	S	0,255	135
ACA	T	0,416	127
ACC	T	0,144	44
ACG	T	0,066	20
ACU	T	0,374	114
GUA	V	0,451	144
GUC	V	0,069	22
GUG	V	0,091	29
GUU	V	0,389	124
UGG	W	1	51
UAC	Y	0,288	132
UAU	Y	0,712	327
UAA	*	0,463	189
UAG	*	0,26	106
UGA	*	0,277	113

PRILOGA D

Stolpični diagram uporabe kodonov *Verticillium nonalfalfae* mitohondrijske DNA
Diagram predstavlja kodone (x-os) in odstotke njihovega pojavljanja (y-os) v *V. nonalfalfae* mitohondrijskem genomu.

Codon usage histogram of the *Verticillium nonalfalfae* mitochondrial DNA
The graph shows codons (x-axis) and percentages of their appearances (y-axis) in the *V. nonalfalfae* mitochondrial genome.



PRILOGA E

orf414 karakterizacija
orf414 characterization

BLASTn ter BLASTx analiza *orf414* zaporedja, e > 10^{-5} z > 50 % pokritostjo

BLASTn: Nucleotide collection (nr/nt)

```
Database: Nucleotide collection (nt)
           33,804,645 sequences; 108,060,332,819 total letters
Query= orf414

Length=789

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

gb|KP994403.1|  Verticillium dahliae isolate F624 tRNA-Lys gen...  1276    0.0
gb|KP994402.1|  Verticillium dahliae isolate F525 tRNA-Lys gen...  1276    0.0
gb|GU291307.1|  Verticillium dahliae isolate V539I cytochrome ...  1276    0.0
gb|GU291306.1|  Verticillium dahliae isolate V396I cytochrome ...  1276    0.0
gb|GU291305.1|  Verticillium dahliae isolate MP89 cytochrome o...  1276    0.0
gb|GU291304.1|  Verticillium dahliae isolate Fca21 cytochrome ...  1276    0.0
gb|KP994401.1|  Verticillium dahliae strain V44 tRNA-Lys gene,...  1270    0.0
gb|GU291303.1|  Verticillium dahliae isolate V137I cytochrome ...  1270    0.0
gb|DQ351941.1|  Verticillium dahliae mitochondrion, complete g...  1267    0.0
gb|CP009079.1|  Verticillium dahliae JR2 chromosome 2, complet...  852     0.0

ALIGNMENTS
>gb|KP994403.1| Verticillium dahliae isolate F624 tRNA-Lys gene, partial sequence;
mitochondrial
Length=1060

 Score = 1276 bits (1414),  Expect = 0.0
 Identities = 755/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1    CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  140  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  199

Query  61   TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
            ||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  200  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  259

Query  121  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
            |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  260  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  319

Query  181  AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
            ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||||
Sbjct  320  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  379

Query  241  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  380  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  439

Query  301  TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
            | |||||||||||  || |||||||||||| ||||||||||||||||||||||||||||||
Sbjct  440  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  499

Query  361  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
            |||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  500  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  559
```

```
Query  421  AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
            |||||||| || |||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  560  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  619

Query  481  CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
             || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  620  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  679

Query  541  AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
            |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  680  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  739

Query  601  TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
            |||| |  ||||| || ||||||||| |||||||||||||||||||||||||||||| |
Sbjct  740  TTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  799

Query  661  ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
            |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  800  ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  859

Query  721  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  860  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  919

Query  781  GTAGTAT  787
            |||||||
Sbjct  920  GTAGTAT  926
```

```
>gb|KP994402.1| Verticillium dahliae isolate F525 tRNA-Lys gene, partial sequence;
mitochondrial
Length=1045

 Score = 1276 bits (1414),  Expect = 0.0
 Identities = 755/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1    CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  125  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  184

Query  61   TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
            |||||||||||||||| |||||||||||||||||||||||||||||| |||||||
Sbjct  185  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  244

Query  121  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
            |||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  245  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  304

Query  181  AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
            ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  305  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  364

Query  241  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  365  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  424

Query  301  TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
            | |||||||||| || |||||||||| ||||||||||||||||||||||||||||||||
Sbjct  425  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  484

Query  361  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
            |||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  485  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  544

Query  421  AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
            |||||||| || |||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  545  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  604

Query  481  CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
             || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  605  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  664
```

```
Query  541  AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
            ||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  665  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  724

Query  601  TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
            |||| |  ||||| || ||||||||| |||||||||||||||||||||||||||||| |
Sbjct  725  TTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  784

Query  661  ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
            |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  785  ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  844

Query  721  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  845  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  904

Query  781  GTAGTAT  787
            |||||||
Sbjct  905  GTAGTAT  911


>gb|GU291307.1| Verticillium dahliae isolate V539I cytochrome oxidase subunit
III (cox3) gene, partial cds; tRNA-Lys, tRNA-Gly, tRNA-Asp,
and tRNA-Ser genes, complete sequence; and NADH dehydrogenase
subunit 6 (nad6) gene, partial cds; mitochondrial
Length=1728

 Score = 1276 bits (1414),  Expect = 0.0
 Identities = 755/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1     CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  265   CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  324

Query  61    TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
             ||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  325   TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  384

Query  121   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
             |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  385   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  444

Query  181   AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
             ||||||||||||||||||  | |||||||||||||||||||||||||||||||||||||
Sbjct  445   AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  504

Query  241   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  505   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  564

Query  301   TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
             | |||||||||||  || |||||||||||| |||||||||||||||||||||||||||||
Sbjct  565   TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  624

Query  361   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
             |||||||||||||||||||||||||||||||||||||   | ||||||||||||||||||
Sbjct  625   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  684

Query  421   AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
             |||||||| || |||||||||||| |||||||||||||||||||||||||||||||| |||
Sbjct  685   AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  744

Query  481   CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
              || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  745   TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  804

Query  541   AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
             ||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct  805   AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  864

Query  601   TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
```

```
            |||| |  |||||  || ||||||||| |||||||||||||||||||||||||||||| |
Sbjct  865  TTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  924

Query  661  ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
            |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  925  ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  984

Query  721  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  985  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  1044

Query  781  GTAGTAT  787
            |||||||
Sbjct  1045 GTAGTAT  1051


>gb|GU291306.1| Verticillium dahliae isolate V396I cytochrome oxidase subunit
III (cox3) gene, partial cds; tRNA-Lys, tRNA-Gly, tRNA-Asp,
and tRNA-Ser genes, complete sequence; and NADH dehydrogenase
subunit 6 (nad6) gene, partial cds; mitochondrial
Length=1754

 Score = 1276 bits (1414),  Expect = 0.0
 Identities = 755/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1    CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  276  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  335

Query  61   TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
            ||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  336  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  395

Query  121  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
            |||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  396  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  455

Query  181  AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
            ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  456  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  515

Query  241  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  516  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  575

Query  301  TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
            | |||||||||||| || |||||||||||| ||||||||||||||||||||||||||||||
Sbjct  576  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  635

Query  361  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
            |||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  636  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  695

Query  421  AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
            |||||||| || |||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  696  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  755

Query  481  CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
             || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  756  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  815

Query  541  AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
            |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  816  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  875

Query  601  TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
            |||| |  |||||  || ||||||||| |||||||||||||||||||||||||||||| |
Sbjct  876  TTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  935

Query  661  ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
            |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  936   ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  995

Query  721   CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  996   CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  1055

Query  781   GTAGTAT  787
              |||||||
Sbjct  1056  GTAGTAT  1062


>gb|GU291305.1| Verticillium dahliae isolate MP89 cytochrome oxidase subunit
III (cox3) gene, partial cds; tRNA-Lys, tRNA-Gly, tRNA-Asp,
and tRNA-Ser genes, complete sequence; and NADH dehydrogenase
subunit 6 (nad6) gene, partial cds; mitochondrial
Length=1662

 Score = 1276 bits (1414),  Expect = 0.0
 Identities = 755/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1     CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  233   CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  292

Query  61    TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              |||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  293   TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  352

Query  121   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  353   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  412

Query  181   AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  413   AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  472

Query  241   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  473   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  532

Query  301   TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||||| || ||||||||||| ||||||||||||||||||||||||||||||
Sbjct  533   TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  592

Query  361   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              |||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  593   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  652

Query  421   AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              |||||||| || |||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  653   AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  712

Query  481   CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  713   TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  772

Query  541   AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
Sbjct  773   AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  832

Query  601   TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
              |||| |  |||||  || ||||||||| |||||||||||||||||||||||||||||| |
Sbjct  833   TTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  892

Query  661   ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
              |||||||||| ||||||||| |||||||||||||||||||||||||||||||||||||||
Sbjct  893   ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  952

Query  721   CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  953   CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  1012
```

```
Query  781   GTAGTAT  787
              |||||||
Sbjct  1013  GTAGTAT  1019


>gb|GU291304.1| Verticillium dahliae isolate Fca21 cytochrome oxidase subunit
III (cox3) gene, partial cds; tRNA-Lys, tRNA-Gly, tRNA-Asp,
and tRNA-Ser genes, complete sequence; and NADH dehydrogenase
subunit 6 (nad6) gene, partial cds; mitochondrial
Length=1644

 Score = 1276 bits (1414),  Expect = 0.0
 Identities = 755/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1    CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  208  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  267

Query  61   TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
            ||||||||||||||| ||||||||||||||||||||||||||||||||| |||||||
Sbjct  268  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  327

Query  121  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
            |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  328  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  387

Query  181  AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
            ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  388  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  447

Query  241  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  448  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  507

Query  301  TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
            | |||||||||||  || |||||||||||| ||||||||||||||||||||||||||||
Sbjct  508  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  567

Query  361  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
            |||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  568  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  627

Query  421  AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
            |||||||| || |||||||||||| ||||||||||||||||||||||||||||| |||
Sbjct  628  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  687

Query  481  CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
             || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  688  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  747

Query  541  AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
            |||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||
Sbjct  748  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  807

Query  601  TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
            |||| |  ||||| || ||||||||| |||||||||||||||||||||||||||||| |
Sbjct  808  TTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  867

Query  661  ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
            |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  868  ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  927

Query  721  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  928  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  987

Query  781  GTAGTAT  787
            |||||||
Sbjct  988  GTAGTAT  994
```

```
>gb|KP994401.1| Verticillium dahliae strain V44 tRNA-Lys gene, partial sequence;
mitochondrial
Length=1060

 Score = 1270 bits (1408),  Expect = 0.0
 Identities = 754/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1    CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  140  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  199

Query  61   TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
            |||||||||||||||| ||||||||||||||||||||||||||||||||| |||||||
Sbjct  200  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  259

Query  121  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
            |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  260  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  319

Query  181  AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
            ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||||
Sbjct  320  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  379

Query  241  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  380  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  439

Query  301  TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
            | |||||||||| || |||||||||||| ||||||||||||||||||||||||||||||
Sbjct  440  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  499

Query  361  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
            |||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  500  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  559

Query  421  AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
            |||||||| || |||||||||||| ||||||||||||||||||||||||||||||| |||
Sbjct  560  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  619

Query  481  CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
             || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  620  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  679

Query  541  AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
            |||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||
Sbjct  680  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  739

Query  601  TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
            |||| |  ||||| || ||||| ||| |||||||||||||||||||||||||||||| |
Sbjct  740  TTGAAGGAGTTGATGATCATCAATCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  799

Query  661  ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
            |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  800  ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  859

Query  721  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  860  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  919

Query  781  GTAGTAT  787
            |||||||
Sbjct  920  GTAGTAT  926


>gb|GU291303.1| Verticillium dahliae isolate V137I cytochrome oxidase subunit
III (cox3) gene, partial cds; tRNA-Lys, tRNA-Gly, tRNA-Asp,
and tRNA-Ser genes, complete sequence; and NADH dehydrogenase
subunit 6 (nad6) gene, partial cds; mitochondrial
Length=1715
```

```
 Score = 1270 bits (1408),  Expect = 0.0
 Identities = 754/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1     CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  267   CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  326

Query  61    TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
             |||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  327   TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  386

Query  121   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
             |||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  387   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  446

Query  181   AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
             ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  447   AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  506

Query  241   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  507   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  566

Query  301   TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
             | |||||||||||| || |||||||||||| ||||||||||||||||||||||||||||||
Sbjct  567   TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  626

Query  361   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
             |||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  627   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  686

Query  421   AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
             |||||||| || |||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  687   AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  746

Query  481   CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
              || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  747   TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  806

Query  541   AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
             |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  807   AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  866

Query  601   TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
             |||| |  |||||  || ||||| ||| |||||||||||||||||||||||||||||| |
Sbjct  867   TTGAAGGAGTTGATGATCATCAATCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  926

Query  661   ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
             |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  927   ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  986

Query  721   CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  987   CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  1046

Query  781   GTAGTAT  787
             |||||||
Sbjct  1047  GTAGTAT  1053


>gb|DQ351941.1| Verticillium dahliae mitochondrion, complete genome
Length=27184

 Score = 1267 bits (1404),  Expect = 0.0
 Identities = 753/787 (96%), Gaps = 0/787 (0%)
 Strand=Plus/Plus

Query  1      CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  15557  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  15616
```

```
Query  61     TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              ||||||||||||||| ||||||||||||||||||||||||||||||||||| |||||||
Sbjct  15617  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  15676

Query  121    CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  15677  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  15736

Query  181    AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||||
Sbjct  15737  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  15796

Query  241    GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  15797  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  15856

Query  301    TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||||  || ||||||||||| ||||||||||||||||||||||||| |||
Sbjct  15857  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGAAATT  15916

Query  361    ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              |||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  15917  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  15976

Query  421    AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              |||||||| || |||||||||||| || |||||||||||||||||||||||| |||
Sbjct  15977  AAGCTAAATACCTAGAAGCAAATAAGTGAATAGAAAAAAATAAACTTGAACTATTAGAAT  16036

Query  481    CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16037  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  16096

Query  541    AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  16097  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  16156

Query  601    TTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACG  660
              |||| |  ||||||  || ||||||||| |||||||||||||||||||||||||||||| |
Sbjct  16157  TTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATG  16216

Query  661    ATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  720
              |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  16217  ATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACTT  16276

Query  721    CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16277  CTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTAG  16336

Query  781    GTAGTAT  787
              |||||||
Sbjct  16337  GTAGTAT  16343


>gb|CP009079.1| Verticillium dahliae JR2 chromosome 2, complete sequence
Length=4277765

 Score = 852 bits (944),  Expect = 0.0
 Identities = 533/573 (93%), Gaps = 3/573 (1%)
 Strand=Plus/Plus

Query  1        CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3466907  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  3466966

Query  61       TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
                ||||||||||||||| ||||||||||||||||||||||||||||||||||| |||||||
Sbjct  3466967  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  3467026

Query  121      CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
                |||||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  3467027  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  3467086
```

```
Query  181      AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
                |||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  3467087  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  3467146

Query  241      GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3467147  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  3467206

Query  301      TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
                | |||||||||||  || |||||||||||| ||||||||||||||||||||||||||||||
Sbjct  3467207  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  3467266

Query  361      ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
                |||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  3467267  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  3467326

Query  421      AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
                |||||||| || |||||||||||| ||||||||||||||||||||||||||||||| |||
Sbjct  3467327  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  3467386

Query  481      CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAA---CAATTTG  537
                 || ||||||||| |||||||||||||||||||| ||||||| ||||| |    |   |
Sbjct  3467387  TTACAATGAATCCAGAAGAACTTAAACGTAAAAGGGAAGATTTTGAAGGAGTTGATAATC  3467446

Query  538      ATGAAATTGAAGCTAAAAAAGTAAAAATGAATC  570
                || ||  | || ||||||| ||||||||| ||||
Sbjct  3467447  ATCAACCTCAAACTAAAAGAGTAAAAATAAATC  3467479


 Score = 174 bits (192),  Expect = 3e-39
 Identities = 113/124 (91%), Gaps = 0/124 (0%)
 Strand=Plus/Plus

Query  563      AATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATTTTGAGGAGGTTGAAAATAATCA  622
                ||||||||| |||||||||||||||||||||||||||||||| |  ||||| ||| ||||
Sbjct  3467391  AATGAATCCAGAAGAACTTAAACGTAAAAGGGAAGATTTTGAAGGAGTTGATAATCATCA  3467450

Query  623      ACCTCCAACTAAAAGAGTAAAAATAAATCATAATAACGATGATAATAATAACGGACAAAG  682
                |||||  |||||||||||||||||||||||||||||  ||| ||||||| ||||||||| |
Sbjct  3467451  ACCTCAAACTAAAAGAGTAAAAATAAATCATAATAATGATAATAATAACAACGGACAAGG  3467510

Query  683      TGGT  686
                ||||
Sbjct  3467511  TGGT  3467514
```

BLASTn: Reference RNA sequences (refseq_rna)

```
No significant similarity found.
```

BLASTn: Reference genomic sequences (refseq_genomic)

```
Database: NCBI Genomic Reference Sequences
          16,767,017 sequences; 729,514,999,410 total letters

Query= orf414

Length=789

                                                                     Score     E
Sequences producing significant alignments:                         (Bits)  Value

ref|NW_009276970.1|  Verticillium dahliae VdLs.17 supercont1.53 m...  1282    0.0
ref|NW_009276969.1|  Verticillium dahliae VdLs.17 supercont1.54 m...  1275    0.0
ref|NC_008248.1|  Verticillium dahliae mitochondrion, complete ge...  1271    0.0
ref|NW_009276924.1|  Verticillium dahliae VdLs.17 supercont1.21 g...  246     3e-060
ref|NW_003315033.1|  Verticillium albo-atrum VaMs.102 supercont1....  80.5    3e-010
```

```
>ref|NW_009276970.1| Verticillium dahliae VdLs.17 supercont1.53 mitochondrial scaffold,
whole genome shotgun sequence
Length=42389

 Score = 1282 bits (694),  Expect = 0.0
 Identities = 757/788 (96%), Gaps = 2/788 (0%)
 Strand=Plus/Minus

Query  1     CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6084  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  6025

Query  61    TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
             ||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  6024  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  5965

Query  121   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
             |||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  5964  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  5905

Query  181   AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
             ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  5904  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  5845

Query  241   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5844  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  5785

Query  301   TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAAGGTTTATATCAAGTAAGACATT  360
             | |||||||||| || |||||||||| ||||||||||||||||||||||||||||||||
Sbjct  5784  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  5725

Query  361   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
             ||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  5724  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  5665

Query  421   AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
             |||||||| || |||||||||||| |||||||||||||||||||||||||||||||| |||
Sbjct  5664  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  5605

Query  481   CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
              || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5604  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  5545

Query  541   AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
             |||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
Sbjct  5544  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  5485

Query  601   TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
             |||| ||| |||||  || ||||||||| ||||||||||||||||||||||||||||||
Sbjct  5484  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  5426

Query  660   GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
             |||||||||| ||||||||| |||||||||||||||||||||||||||||||||||||
Sbjct  5425  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  5366

Query  720   TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5365  TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  5306

Query  780   GGTAGTAT  787
             ||||||||
Sbjct  5305  GGTAGTAT  5298


 Score = 1275 bits (690),  Expect = 0.0
 Identities = 756/788 (96%), Gaps = 3/788 (0%)
 Strand=Plus/Minus

Query  1      CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  33051  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  32992
```

```
Query  61     TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              ||||||||||||||| |||||||||||||||||||||||||||||||||| |||||||
Sbjct  32991  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  32932

Query  121    CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  32931  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  32872

Query  181    AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  32871  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  32812

Query  241    GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  32811  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  32752

Query  301    TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||||  || ||||||||||| ||||||||||||||||||||||||||||||
Sbjct  32751  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  32692

Query  361    ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              |||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  32691  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  32632

Query  421    AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              |||||||| || |||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  32631  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  32572

Query  481    CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  32571  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  32512

Query  541    AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct  32511  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  32452

Query  601    TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
              |||| ||| |||||  || ||||||||| ||||||||||||||||||||||||||||||
Sbjct  32451  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  32393

Query  660    GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
              |||||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  32392  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  32333

Query  720    TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
              ||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  32332  TCT-CAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  32274

Query  780    GGTAGTAT  787
              ||||||||
Sbjct  32273  GGTAGTAT  32266


>ref|NW_009276969.1| Verticillium dahliae VdLs.17 supercont1.54 mitochondrial scaffold,
whole genome shotgun sequence
Length=19770

 Score = 1275 bits (690),  Expect = 0.0
 Identities = 756/788 (96%), Gaps = 4/788 (1%)
 Strand=Plus/Plus

Query  1      CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13417  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  13476

Query  61     TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              ||||||||||||||| |||||||||||||||||||||||||||||||||| |||||||
Sbjct  13477  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGG-GATGAGA  13535

Query  121    CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
```

```
Sbjct  13536  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  13595

Query  181    AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||||
Sbjct  13596  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  13655

Query  241    GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13656  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  13715

Query  301    TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||||  || ||||||||||| ||||||||||||||||||||||||||||||
Sbjct  13716  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  13775

Query  361    ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              ||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  13776  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  13835

Query  421    AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              |||||||| || |||||||||||| ||||||||||||||||||||||||||||| |||
Sbjct  13836  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  13895

Query  481    CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13896  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  13955

Query  541    AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  13956  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  14015

Query  601    TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
              |||| ||| |||||  || ||||||||| ||||||||||||||||||||||||||||||
Sbjct  14016  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  14074

Query  660    GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
              |||||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  14075  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  14134

Query  720    TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
              |||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  14135  TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTA-TTTATGATTA  14193

Query  780    GGTAGTAT  787
              ||||||||
Sbjct  14194  GGTAGTAT  14201


>ref|NC_008248.1| Verticillium dahliae mitochondrion, complete genome
Length=27184

 Score = 1271 bits (688),  Expect = 0.0
 Identities = 755/788 (96%), Gaps = 2/788 (0%)
 Strand=Plus/Plus

Query  1      CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  15557  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  15616

Query  61     TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              ||||||||||||||| |||||||||||||||||||||||||||||||||| |||||||
Sbjct  15617  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  15676

Query  121    CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  15677  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  15736

Query  181    AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||||
Sbjct  15737  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  15796

Query  241    GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  15797  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  15856

Query  301    TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||||  || |||||||||||  ||||||||||||||||||||||||||| |||
Sbjct  15857  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGAAATT  15916

Query  361    ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              |||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  15917  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  15976

Query  421    AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              ||||||||  || ||||||||||||  || ||||||||||||||||||||||||| |||
Sbjct  15977  AAGCTAAATACCTAGAAGCAAATAAGTGAATAGAAAAAAATAAACTTGAACTATTAGAAT  16036

Query  481    CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16037  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  16096

Query  541    AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  16097  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  16156

Query  601    TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
              |||| ||| |||||  || |||||||| ||||||||||||||||||||||||||||||
Sbjct  16157  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  16215

Query  660    GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
              |||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  16216  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  16275

Query  720    TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16276  TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  16335

Query  780    GGTAGTAT  787
              ||||||||
Sbjct  16336  GGTAGTAT  16343


>ref|NW_009276924.1| Verticillium dahliae VdLs.17 supercont1.21 genomic scaffold,
whole genome shotgun sequence
Length=641803

 Score = 246 bits (133),  Expect = 3e-060
 Identities = 186/212 (88%), Gaps = 2/212 (1%)
 Strand=Plus/Plus

Query  7       TCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAGTAGACT  66
               |||| || |||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  206871  TCGACCATGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAGTACACT  206930

Query  67      CAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGACTTCAG  126
               || ||||||| |||||||| |||||||| |||||||||||||||| ||  ||||||||
Sbjct  206931  CAAGTCTAGATAGCGATGAGCAAGTTCAGAGTAGTGGATCCGAGGGGGAGCAGACTTCAC  206990

Query  127     AAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAGAAAGAT  186
               | | ||| ||| | |||||||||||| ||||||||| ||||||||| |||| ||||| ||
Sbjct  206991  ACAGTGAGTCATACATTGAAGGTATACATGAGGTTTGTGATTTTGAGGTTG-AGAAACAT  207049

Query  187     CAGGTGATAAACTCACAACTGAAAGATTAGCT  218
               ||||||| |||||  |||||||| ||| ||||
Sbjct  207050  CAGGTGACAAACTTGCAACTGAA-GATAAGCT  207080


>ref|NW_003315033.1| Verticillium albo-atrum VaMs.102 supercont1.6 genomic scaffold,
whole genome shotgun sequence
Length=2315232

 Score = 80.5 bits (43),  Expect = 3e-010
 Identities = 49/52 (94%), Gaps = 0/52 (0%)
 Strand=Plus/Minus
```

```
Query  1        CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGA  52
                |||||||| ||  ||||||||||||||||||||||||||||||||||||||
Sbjct  1659624  CAGAACTCAAATCAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGA  1659573
```

BLASTn: Refseq Representative genomes (refseq_representative_genomes)

```
Database: RefSeq Representative Genome Database
          10,990,398 sequences; 422,001,285,913 total letters


Query= orf414

Length=789

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

ref|NW_009276970.1|  Verticillium dahliae VdLs.17 supercont1.53 m...  1282    0.0
ref|NW_009276969.1|  Verticillium dahliae VdLs.17 supercont1.54 m...  1275    0.0
ref|NW_009276924.1|  Verticillium dahliae VdLs.17 supercont1.21 g...  246     2e-060
ref|NW_003315033.1|  Verticillium albo-atrum VaMs.102 supercont1....  80.5    2e-010


>ref|NW_009276970.1| Verticillium dahliae VdLs.17 supercont1.53 mitochondrial scaffold,
whole genome shotgun sequence
Length=42389

 Score = 1282 bits (694),  Expect = 0.0
 Identities = 757/788 (96%), Gaps = 2/788 (0%)
 Strand=Plus/Minus

Query  1     CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6084  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  6025

Query  61    TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              |||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  6024  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  5965

Query  121   CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  5964  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  5905

Query  181   AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  5904  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  5845

Query  241   GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5844  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  5785

Query  301   TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||||  || |||||||||||  ||||||||||||||||||||||||||||||
Sbjct  5784  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  5725

Query  361   ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              |||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  5724  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  5665

Query  421   AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              |||||||| || |||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  5664  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  5605

Query  481   CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5604  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  5545

Query  541   AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  5544  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  5485
```

```
Query  601   TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
             ||||  |||  |||||  ||  ||||||||| ||||||||||||||||||||||||||||||
Sbjct  5484  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  5426

Query  660   GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
             ||||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  5425  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  5366

Query  720   TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5365  TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  5306

Query  780   GGTAGTAT  787
             ||||||||
Sbjct  5305  GGTAGTAT  5298


 Score = 1275 bits (690),  Expect = 0.0
 Identities = 756/788 (96%), Gaps = 3/788 (0%)
 Strand=Plus/Minus

Query  1      CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  33051  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  32992

Query  61     TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
             |||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  32991  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  32932

Query  121    CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
             |||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  32931  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  32872

Query  181    AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
             ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||||
Sbjct  32871  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  32812

Query  241    GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  32811  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  32752

Query  301    TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
             | |||||||||||  || ||||||||||| ||||||||||||||||||||||||||||||
Sbjct  32751  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  32692

Query  361    ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
             |||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  32691  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  32632

Query  421    AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
             |||||||| || |||||||||||| ||||||||||||||||||||||||||||| |||
Sbjct  32631  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  32572

Query  481    CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
              || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  32571  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  32512

Query  541    AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
             |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  32511  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  32452

Query  601    TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
             ||||  |||  |||||  ||  ||||||||| ||||||||||||||||||||||||||||
Sbjct  32451  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  32393

Query  660    GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
             |||||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  32392  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  32333

Query  720    TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
             ||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  32332  TCT-CAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  32274

Query  780    GGTAGTAT  787
              ||||||||
Sbjct  32273  GGTAGTAT  32266


>ref|NW_009276969.1| Verticillium dahliae VdLs.17 supercont1.54 mitochondrial scaffold,
whole genome shotgun sequence
Length=19770

 Score = 1275 bits (690),  Expect = 0.0
 Identities = 756/788 (96%), Gaps = 4/788 (1%)
 Strand=Plus/Plus

Query  1      CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13417  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  13476

Query  61     TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              |||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  13477  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGG-GATGAGA  13535

Query  121    CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
Sbjct  13536  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  13595

Query  181    AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  13596  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  13655

Query  241    GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13656  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  13715

Query  301    TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||| || |||||||||| ||||||||||||||||||||||||||||||
Sbjct  13716  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGACATT  13775

Query  361    ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              |||||||||||||||||||||||||||||||||||  | ||||||||||||||||||
Sbjct  13776  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  13835

Query  421    AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              |||||||| || |||||||||| |||||||||||||||||||||||||||||| |||
Sbjct  13836  AAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTTGAACTATTAGAAT  13895

Query  481    CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13896  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  13955

Query  541    AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  13956  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  14015

Query  601    TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
              |||| ||| |||||  || ||||||||| ||||||||||||||||||||||||||||||
Sbjct  14016  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  14074

Query  660    GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
              |||||||||| ||||||||| |||||||||||||||||||||||||||||||||||||
Sbjct  14075  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  14134

Query  720    TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
              |||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  14135  TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTA-TTTATGATTA  14193

Query  780    GGTAGTAT  787
              ||||||||
Sbjct  14194  GGTAGTAT  14201
```

```
>ref|NW_009276924.1| Verticillium dahliae VdLs.17 supercont1.21 genomic scaffold,
whole genome shotgun sequence
Length=641803

 Score = 246 bits (133),  Expect = 2e-060
 Identities = 186/212 (88%), Gaps = 2/212 (1%)
 Strand=Plus/Plus

Query  7       TCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAGTAGACT  66
               |||| || ||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  206871  TCGACCATGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAGTACACT  206930

Query  67      CAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGACTTCAG  126
               || ||||||| |||||||| |||||||| |||||||||||||||| ||  ||||||||
Sbjct  206931  CAAGTCTAGATAGCGATGAGCAAGTTCAGAGTAGTGGATCCGAGGGGGAGCAGACTTCAC  206990

Query  127     AAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAGAAAGAT  186
               | | ||| ||| | ||||||||||| ||||||||| ||||||||| |||| ||||| ||
Sbjct  206991  ACAGTGAGTCATACATTGAAGGTATACATGAGGTTTGTGATTTTGAGGTTG-AGAAACAT  207049

Query  187     CAGGTGATAAACTCACAACTGAAAGATTAGCT  218
               ||||||| |||||  |||||||| ||| ||||
Sbjct  207050  CAGGTGACAAACTTGCAACTGAA-GATAAGCT  207080


>ref|NW_003315033.1| Verticillium albo-atrum VaMs.102 supercont1.6 genomic scaffold,
whole genome shotgun sequence
Length=2315232

 Score = 80.5 bits (43),  Expect = 2e-010
 Identities = 49/52 (94%), Gaps = 0/52 (0%)
 Strand=Plus/Minus

Query  1        CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGA  52
                |||||||| ||  ||||||||||||||||||||||||||||||||||||||
Sbjct  1659624  CAGAACTCAAATCAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGA  1659573
```

BLASTn: NCBI Genomes (chromsome)

```
Database: NCBI Chromosome Sequences
          86,835 sequences; 362,426,053,891 total letters

Query= orf414

Length=789

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

ref|NC_008248.1|  Verticillium dahliae mitochondrion, complete ge...  1271    0.0
ref|NW_003315033.1|  Verticillium albo-atrum VaMs.102 supercont1....  80.5    2e-010


>ref|NC_008248.1| Verticillium dahliae mitochondrion, complete genome
Length=27184

 Score = 1271 bits (688),  Expect = 0.0
 Identities = 755/788 (96%), Gaps = 2/788 (0%)
 Strand=Plus/Plus

Query  1      CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  15557  CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGACCCTGAAG  15616

Query  61     TAGACTCAGGTCTAGACAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGAGATGAGA  120
              ||||||||||||||| |||||||||||||||||||||||||||||||| |||||||
Sbjct  15617  TAGACTCAGGTCTAGATAGCGATGAACAAGTTCAAAGTAGTGGATCCGAGGGGGATGAGA  15676

Query  121    CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTCTGATTTTGAAGTTGAAG  180
              |||||||||||||||||||||||||||||||||||||||||| ||||||||| |||||||
```

BLASTn: NCBI Genomes (chromsome)

```
Sbjct  15677  CTTCAGAAAATGAATCAGAAATTGAAGGTATAAATGAGGTTTGTGATTTTGAGGTTGAAG  15736

Query  181    AAAGATCAGGTGATAAACTCACAACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  240
              ||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||||
Sbjct  15737  AAAGATCAGGTGATAAACTTGCTACTGAAAGATTAGCTAATGATCAGACTCATCTTCTTA  15796

Query  241    GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  15797  GAGCTTTAAGACATGGAGAACAAGCTTCTATAGATAAAATACAAGAGAGATACCCTGCTT  15856

Query  301    TCTTTGATGAAGGTAGTGGAAATTCTTCCATAAAAAAAGGTTTATATCAAGTAAGACATT  360
              | |||||||||||  || ||||||||||| ||||||||||||||||||||||||| |||
Sbjct  15857  TTTTTGATGAAGGGAGCGGAAATTCTTCCGTAAAAAAAGGTTTATATCAAGTAAGAAATT  15916

Query  361    ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAAATAGATAGAGAAG  420
              ||||||||||||||||||||||||||||||||||||  | ||||||||||||||||||||
Sbjct  15917  ATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACCTTGAAAGAAATAGATAGAGAAG  15976

Query  421    AAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTTGAACTATTTGAAC  480
              |||||||| || |||||||||||| || ||||||||||||||||||||||||| |||
Sbjct  15977  AAGCTAAATACCTAGAAGCAAATAAGTGAATAGAAAAAAATAAACTTGAACTATTAGAAT  16036

Query  481    CTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  540
               || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16037  TTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAAGAACAATTTGATG  16096

Query  541    AAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGTAAAAGGGAAGATT  600
              |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  16097  AAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGTAAAAGGGAAGATT  16156

Query  601    TTGA-GGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATAAATCATAATAAC  659
              |||| ||| |||||  || ||||||||| ||||||||||||||||||||||||||||||
Sbjct  16157  TTGAAGGA-GTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATAAATCATAATAAT  16215

Query  660    GATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  719
              |||||||||||| ||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  16216  GATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATTTCATCTGGTACT  16275

Query  720    TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  779
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16276  TCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATTATTTTATGATTA  16335

Query  780    GGTAGTAT  787
              ||||||||
Sbjct  16336  GGTAGTAT  16343


>ref|NW_003315033.1| Verticillium albo-atrum VaMs.102 supercont1.6 genomic scaffold,
whole genome shotgun sequence
Length=2315232

 Score = 80.5 bits (43),  Expect = 2e-010
 Identities = 49/52 (94%), Gaps = 0/52 (0%)
 Strand=Plus/Minus

Query  1        CAGAACTCGAACAAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGA  52
                ||||||||  ||  ||||||||||||||||||||||||||||||||||||||
Sbjct  1659624  CAGAACTCAAATCAGATGATCATGATTCTCATGATGATTCTTTAGCTACTGA  1659573
```

BLASTn: Expressed sequence tags (est)

```
Database: Database of GenBank+EMBL+DDBJ sequences from EST Divisions
          76,041,847 sequences; 42,369,095,621 total letters
Query= orf414

Length=789

                                                                  Score     E
```

```
Sequences producing significant alignments:                          (Bits)  Value

gb|GR299985.1|  G969P31RA11.T0 Phytophthora infestans T30-4 NO...  686     0.0

ALIGNMENTS
>gb|GR299985.1| G969P31RA11.T0 Phytophthora infestans T30-4 NORMALIZED Phytophthora
infestans cDNA, mRNA sequence.
Length=705

 Score = 686 bits (760),  Expect = 0.0
 Identities = 416/440 (95%), Gaps = 0/440 (0%)
 Strand=Plus/Minus

Query  348  CAAGTAAGACATTATATAGAAGAAGAATTTGATCTTGAAGAATTAGAAACACTAAAAGAA  407
            |||||||| ||||||||||||||||||||||||||||||| ||||||||  | ||||||
Sbjct  705  CAAGTAAGCCATTATATAGAAGAAGAATTTGATCTTGAAGAGTTAGAAACCTTGAAAGAA  646

Query  408  ATAGATAGAGAAGAAGCTAAACACATAGAAGCAAATAGGTTAATAGAAAAAAATAAACTT  467
            |||||||||||||||||||| || |||||||||||| ||||||||||||||||||||||
Sbjct  645  ATAGATAGAGAAGAAGCTAAATACCTAGAAGCAAATAAGTTAATAGAAAAAAATAAACTT  586

Query  468  GAACTATTTGAACCTATAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAA  527
            |||||||| |||  || ||||||||||||||||||||||||||||||||||||||||||
Sbjct  585  GAACTATTAGAATTTACAATGAATCCTGAAGAACTTAAACGTAAAAGAGAAGATTCTGAA  526

Query  528  GAACAATTTGATGAAATTGAAGCTAAAAAAGTAAAAATGAATCCTGAAGAACTTAAACGT  587
            |||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  525  GAACAATTTGATGAAATTGAAGCTAAAAAAGTAAAAATGAACCCTGAAGAACTTAAACGT  466

Query  588  AAAAGGGAAGATTTTGAGGAGGTTGAAAATAATCAACCTCCAACTAAAAGAGTAAAAATA  647
            |||||||||||||||| |  |||||  || ||||||||| ||||||||||||||||||||
Sbjct  465  AAAAGGGAAGATTTTGAAGGAGTTGATGATCATCAACCTCAAACTAAAAGAGTAAAAATA  406

Query  648  AATCATAATAACGATGATAATAATAACGGACAAAGTGGTATAGGCCCTTGTTCTGGTATT  707
            |||||||||| |||||||||| ||||||||| |||||||||||||||||||||||||||
Sbjct  405  AATCATAATAATGATGATAATAACAACGGACAAGGTGGTATAGGCCCTTGTTCTGGTATT  346

Query  708  TCATCTGGTACTTCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATAACTACTTTAATT  767
            |||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  345  TCATCTGGTACTTCTTCAGAAGAAACTTCTACTAATGCGCGTAGTATACCTACTTTAATT  286

Query  768  ATTTTATGATTAGGTAGTAT  787
            ||||||||||||||||||||
Sbjct  285  ATTTTATGATTAGGTAGTAT  266
```

BLASTn: Genomic survey sequences (gss)

```
No significant similarity found.
```

BLASTn: High throughput genomic sequences (HTGS)

```
No significant similarity found.
```

BLASTn: Transcriptome Shotgun Assembly (TSA) sequences (tsa_all)

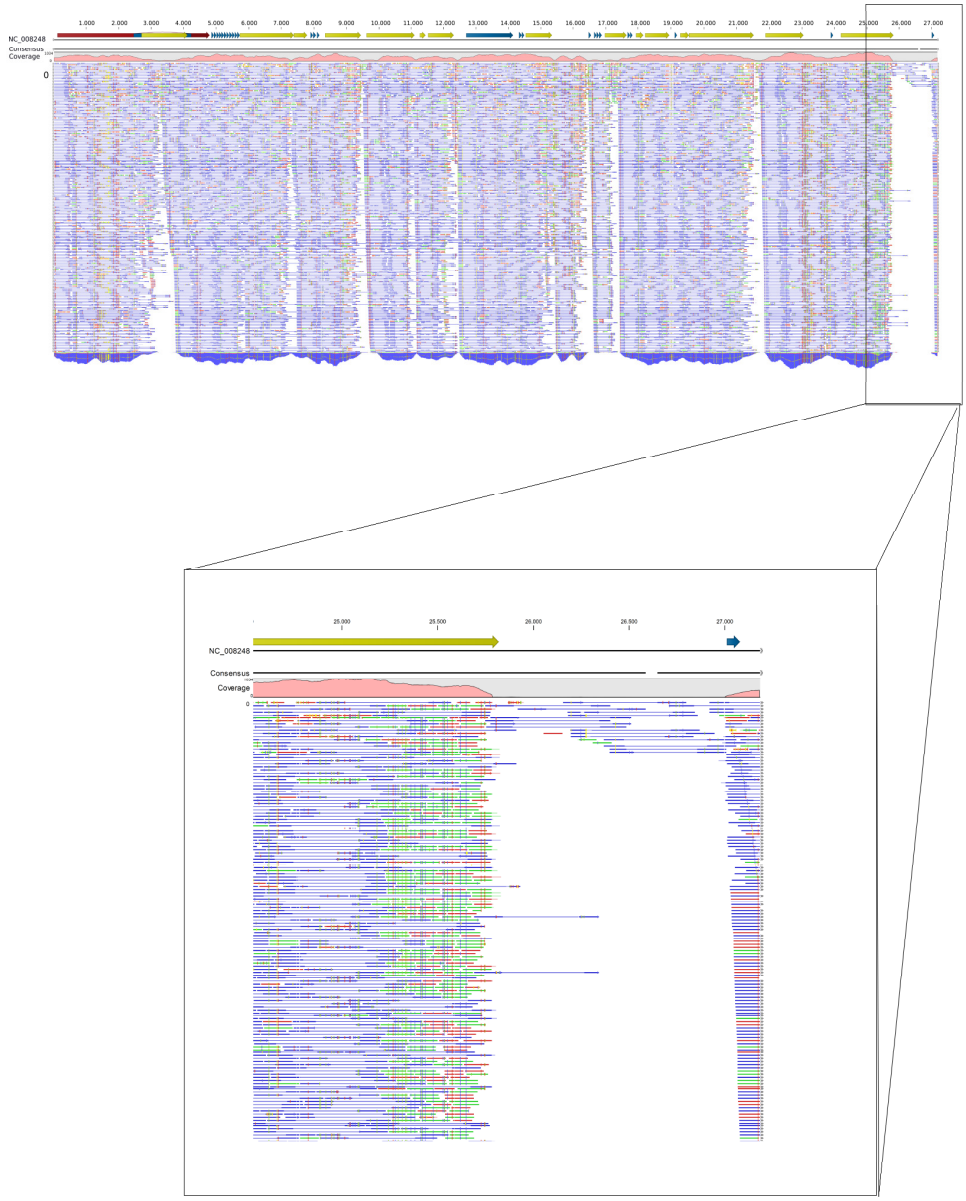
```
No significant similarity found.
```

BLASTx: Non-redundant protein sequences (nr)

```
No significant similarity found.
```

PRILOGA F

Slika kartiranja dolžinskega polimorfizma
Figure of length polymorphism mapping



PRILOGA G

Rezultati pomnoževanja dolžinskega polimorfizma pri 96-ih vzorcih gliv iz rodu Verticillium
Length polymorphism PCR amplification results of 96 *Verticillium* fungal samples

	V. alboatrum	*V. alfalfae*	*V. dahliae*	*V. isaacii*	*V. longisporum*	*V. nonalfalfae*	*V. nigrescens*	*V. nubilum*	*V. tricorpus*
1400	0	0	22	0	0	0	0	1	0
400	0	5	0	0	2	44	0	0	0
Neuspelo	6	0	4	3	0	5	2	0	2
Analizirano	6	5	26	3	2	49	2	1	2

PRILOGA H

Geni s Ka/Ks koeficientom večjim od 1, ki nakazujejo pozitivno selekcijo
Genes with the Ka/Ks coefficient bigger than 1, which exhibit positive selection

Gen	Ka	Ks	Ka/Ks
chr-unplaced.127-EEY23718.1	0,014	0,004	3,338
chr-unplaced.137-EEY23728.1	0,017	0,015	1,116
chr-unplaced.142-EEY23735.1	0,066	0,060	1,112
chr-unplaced.174-EEY19072.1	0,034	0,032	1,056
chr-unplaced.239_CP009075.1-2328486-2329660	0,095	0,091	1,047
chr-unplaced.244-EEY24045.1	0,033	0,026	1,244
chr-unplaced.258_CP009078.1-3256235-3256432	0,091	0,067	1,361
chr-unplaced.269-EEY22764.1	0,238	0,230	1,031
chr-unplaced.282_CP009080.1-3510391-3510862	0,035	0,031	1,125
chr-unplaced.342_CP009078.1-2207475-2207769	0,090	0,089	1,008
chr-unplaced.365-EEY24078.1	0,047	0,020	2,293
chr-unplaced.369_CP009078.1-660823-662981	0,050	0,032	1,560
chr-unplaced.371_CP009078.1-659494-659874	0,051	0,022	2,287
chr-unplaced.373-EEY22881.1	0,011	0,005	1,996
chr-unplaced.427_CP009077.1-3432280-3432593	0,024	0,019	1,235
chr-unplaced.43-EEY14332.1	0,010	0,007	1,524
chr-unplaced.437-EEY22868.1	0,012	0,006	1,903
chr-unplaced.468_CP009079.1-2606478-2606642	0,044	0,025	1,787
chr-unplaced.492-EEY22875.1	0,073	0,064	1,140
chr1_1435YY.1049-EEY21124.1	0,020	0,010	1,931
chr1_1435YY.1064_CP009082.1-994593-995712	0,056	0,051	1,096
chr1_1435YY.1064-EEY21144.1	0,064	0,060	1,075
chr1_1435YY.11-EEY16413.1	0,279	0,279	1,001
chr1_1435YY.1195_CP009082.1-514465-514871	0,080	0,060	1,330
chr1_1435YY.1195_CP009082.1-514465-514871	0,075	0,070	1,070
chr1_1435YY.1195-EEY21290.1	0,037	0,016	2,276
chr1_1435YY.1214_CP009082.1-420186-420394	0,076	0,042	1,802
chr1_1435YY.1256-EEY23779.1	0,053	0,038	1,399
chr1_1435YY.1305-EEY23839.1	0,016	0,012	1,352
chr1_1435YY.1327-EEY23865.1	0,051	0,036	1,397
chr1_1435YY.23-EEY16428.1	0,016	0,010	1,507
chr1_1435YY.269-EEY16715.1	0,156	0,154	1,014
chr1_1435YY.382-EEY16818.1	0,133	0,121	1,100
chr1_1435YY.494-EEY16937.1	0,027	0,013	2,107
chr1_1435YY.78-EEY16489.1	0,047	0,022	2,167
chr1_1435YY.923-EGY15569.1	0,287	0,252	1,140
chr1_1435YY.947_CP009082.1-1425055-1425481	0,085	0,081	1,051

se nadaljuje

nadaljevanje priloge H

Gen	Ka	Ks	Ka/Ks
chr1_1435YY.98-EEY16512.1	0,138	0,124	1,116
chr1_1435YY.985-EEY21054.1	0,015	0,014	1,069
chr10_684NY.123_CP009076.1-2734571-2735334	0,107	0,088	1,211
chr10_684NY.164-EEY23374.1	0,069	0,068	1,011
chr10_684NY.215-EEY23429.1	0,164	0,138	1,182
chr10_684NY.22_CP009076.1-2147633-2150909	0,048	0,019	2,476
chr10_684NY.267_CP009076.1-3226861-3227196	0,039	0,036	1,094
chr10_684NY.328-EEY23554.1	0,013	0,012	1,067
chr10_684NY.63-EEY23261.1	0,007	0,001	5,930
chr10_684NY.76_CP009076.1-2530092-2530540	0,137	0,124	1,108
chr2_1770NN.1016-EEY17519.1	0,017	0,010	1,638
chr2_1770NN.1051_CP009075.1-724463-724809	0,070	0,062	1,127
chr2_1770NN.1156-EEY18760.1	0,177	0,166	1,067
chr2_1770NN.126-EEY14243.1	0,030	0,028	1,070
chr2_1770NN.488-EEY18151.1	0,015	0,010	1,435
chr2_1770NN.511-EEY18124.1	0,300	0,293	1,023
chr2_1770NN.577_CP009075.1-2416175-2417401	0,062	0,025	2,462
chr2_1770NN.61_CP009075.1-4714113-4714500	0,091	0,085	1,071
chr2_1770NN.706-EEY17861.1	0,025	0,013	1,961
chr2_1770NN.784-EEY17776.1	0,065	0,055	1,176
chr2_1770NN.806_CP009075.1-1558985-1559330	0,070	0,055	1,271
chr2_1770NN.814-EEY17740.1	0,026	0,019	1,316
chr2_1770NN.930-EEY17611.1	0,048	0,021	2,289
chr3_121598YY.1010_CP009080.1-204824-205526	0,024	0,018	1,325
chr3_121598YY.1018-EEY19666.1	0,058	0,038	1,549
chr3_121598YY.1023_CP009080.1-155845-156615	0,098	0,095	1,035
chr3_121598YY.145-EEY21780.1	0,034	0,028	1,210
chr3_121598YY.154-EEY21789.1	0,144	0,105	1,369
chr3_121598YY.188-EEY21824.1	0,486	0,441	1,103
chr3_121598YY.351_CP009077.1-2857860-2858495	0,100	0,096	1,036
chr3_121598YY.351-EGY17935.1	0,594	0,485	1,223
chr3_121598YY.381-EEY22044.1	0,029	0,027	1,097
chr3_121598YY.411-EEY22081.1	0,024	0,023	1,031
chr3_121598YY.420-EEY22090.1	0,010	0,010	1,036
chr3_121598YY.433-EEY22108.1	0,068	0,065	1,047
chr3_121598YY.465_CP009077.1-2242856-2243029	0,056	0,037	1,501
chr3_121598YY.475-EEY19024.1	0,048	0,038	1,275
chr3_121598YY.499-EEY19055.1	0,205	0,186	1,103
chr3_121598YY.627-EEY19210.1	0,019	0,015	1,307
chr3_121598YY.66-EEY23955.1	0,104	0,080	1,303

se nadaljuje

nadaljevanje priloge H

Gen	Ka	Ks	Ka/Ks
chr3_121598YY.674-EEY19263.1	0,111	0,099	1,123
chr3_121598YY.735-EEY19342.1	0,030	0,028	1,064
chr3_121598YY.753-EEY19360.1	0,027	0,026	1,064
chr3_121598YY.829-EEY19444.1	0,011	0,008	1,381
chr3_121598YY.871-EEY19490.1	0,100	0,100	1,009
chr3_121598YY.946-EEY19586.1	0,070	0,067	1,056
chr4_116292YY.108_CP009077.1-415665-415836	0,068	0,023	2,943
chr4_116292YY.13-EEY18959.1	0,054	0,048	1,127
chr4_116292YY.202-EEY18702.1	0,115	0,104	1,100
chr4_116292YY.217_CP009077.1-924875-925011	0,086	0,064	1,348
chr4_116292YY.252-EEY18649.1	0,053	0,050	1,065
chr4_116292YY.274_CP009077.1-1094814-1095914	0,025	0,017	1,479
chr4_116292YY.336-EEY18556.1	0,026	0,026	1,013
chr4_116292YY.44-EEY18921.1	1,958	1,473	1,329
chr4_116292YY.506_CP009077.1-2019987-2020568	0,059	0,034	1,726
chr4_116292YY.547-EEY18280.1	0,009	0,006	1,401
chr4_116292YY.602-EEY23204.1	0,023	0,009	2,485
chr4_116292YY.676-EEY23113.1	0,180	0,128	1,402
chr4_116292YY.709-EEY23063.1	0,007	0,004	1,588
chr4_116292YY.805_CP009078.1-649565-649902	0,041	0,030	1,348
chr5_1271NY.1_CP009076.1-2114605-2115341	0,058	0,027	2,141
chr5_1271NY.149_CP009079.1-1943424-1944015	0,050	0,048	1,047
chr5_1271NY.215-EEY15601.1	0,115	0,094	1,229
chr5_1271NY.257-EEY15648.1	0,038	0,037	1,052
chr5_1271NY.306-EEY15707.1	0,098	0,097	1,005
chr5_1271NY.323_CP009079.1-2614200-2614472	0,023	0,019	1,229
chr5_1271NY.355_CP009079.1-2730553-2730879	0,115	0,080	1,443
chr5_1271NY.41-EEY15426.1	0,015	0,015	1,003
chr5_1271NY.719-EEY16153.1	0,010	0,009	1,215
chr5_1271NY.727-EEY16160.1	0,062	0,034	1,825
chr5_1271NY.77-EEY15465.1	0,024	0,014	1,679
chr5_1271NY.84-EEY15472.1	0,067	0,046	1,461
chr5_1271NY.873-EEY16324.1	0,094	0,092	1,027
chr6_1624YN.135-EEY20511.1	0,025	0,021	1,158
chr6_1624YN.235-EEY20632.1	0,259	0,248	1,045
chr6_1624YN.274-EEY20673.1	0,021	0,011	1,830
chr6_1624YN.304-EEY20734.1	0,305	0,302	1,009
chr6_1624YN.558-EEY22653.1	0,293	0,288	1,017
chr6_1624YN.615-EEY22750.1	0,009	0,008	1,086
chr6_1624YN.666-EEY22827.1	0,028	0,028	1,013

se nadaljuje

nadaljevanje priloge H

Gen	Ka	Ks	Ka/Ks
chr6_1624YN.73-EEY21624.1	0,132	0,116	1,140
chr7_742YN.211-EEY15165.1	0,033	0,029	1,140
chr7_742YN.244-EEY15126.1	0,114	0,109	1,044
chr7_742YN.315-EEY15040.1	0,172	0,167	1,030
chr7_742YN.325-EEY15031.1	0,018	0,014	1,297
chr7_742YN.412-EEY14933.1	0,280	0,215	1,302
chr7_742YN.431-EEY14913.1	0,033	0,031	1,069
chr7_742YN.450_CP009075.1-6470616-6471132	0,095	0,059	1,607
chr7_742YN.467-EEY14875.1	0,719	0,677	1,062
chr7_742YN.498-EEY14844.1	0,012	0,005	2,453
chr7_742YN.538_CP009075.1-6137127-6149939	0,044	0,007	6,467
chr7_742YN.77-EEY15301.1	0,178	0,139	1,279
chr7_742YN.847-EEY14450.1	0,057	0,056	1,020
chr8_117813YY.108-EEY22390.1	0,095	0,084	1,139
chr8_117813YY.133_CP009081.1-454621-455062	0,025	0,023	1,092
chr8_117813YY.282-EEY22196.1	0,010	0,009	1,169
chr8_117813YY.308-EEY22168.1	0,029	0,022	1,281
chr8_117813YY.416-EEY19817.1	0,031	0,027	1,131
chr8_117813YY.575-EEY19996.1	0,019	0,015	1,292
chr8_117813YY.769-EEY20215.1	0,036	0,028	1,280
chr9_120842YY.134-EEY20453.1	0,126	0,096	1,307
chr9_120842YY.177-EEY24010.1	0,015	0,005	2,912
chr9_120842YY.307-EEY21443.1	0,162	0,154	1,052
chr9_120842YY.339-EEY21409.1	0,227	0,183	1,242
chr9_120842YY.518-EEY23665.1	0,030	0,029	1,041

PRILOGA I

Preverjanje kakovosti glede na pojavitve dvoumnih baz po celotni dolžini odčitkov
Quality control of ambiguous base occurences along the read length

1953

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
1	13.584.515	9	0,00
2	13.584.515	7	0,00
3	13.584.515	6	0,00
14	13.584.515	1	0,00
16	13.584.515	5	0,00
17	13.584.515	34	0,00
18	13.584.515	326	0,00
19	13.584.515	79	0,00
20	13.584.515	9	0,00
21	13.584.515	296	0,00
22	13.584.515	1	0,00
23	13.584.515	3	0,00
24	13.584.515	14	0,00
25	13.584.515	7,165	0,05
29	13.584.515	2	0,00
31	13.566.980	1	0,00
34	13.533.611	1	0,00
35	13.513.750	7	0,00
36	13.488.407	358	0,00
37	13.477.091	7	0,00
38	13.461.628	10	0,00
39	13.440.977	27	0,00
40	13.415.499	12	0,00
41	13.382.685	18	0,00
42	13.367.665	4	0,00
43	13.348.142	6	0,00
44	13.321.793	2	0,00
45	13.285.277	1	0,00
46	13.236.026	3	0,00
47	13.214.711	1	0,00
48	13.186.260	56	0,00

se nadaljuje

nadaljevanje priloge I - 1953

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
51	13.024.356	5	0,00
52	12.995.080	2	0,00
53	12.956.196	8	0,00
54	12.903.410	3	0,00
55	12.831.586	46	0,00
56	12.733.340	9,033	0,07
57	12.694.909	110	0,00
58	12.643.187	728	0,01
59	12.575.128	50	0,00
61	12.362.064	1	0,00
64	12.158.135	254	0,00
65	12.052.006	6	0,00
66	11.912.894	4	0,00
67	11.857.366	9	0,00
68	11.783.264	55	0,00
69	11.682.982	24	0,00
70	11.554.564	19	0,00
72	11.313.438	383	0,00
73	11.221.330	491	0,00
74	11.099.723	16	0,00
77	10.670.485	5	0,00
79	10.424.575	490	0,00
80	10.251.272	1,247	0,01
81	10.029.502	1	0,00
82	9.943.579	31	0,00
86	9.284.082	4	0,00
87	9.187.160	3	0,00
88	9.071.394	2	0,00
89	8.922.978	42	0,00
90	8.740.679	2	0,00
91	8.514.362	8	0,00
92	8.422.472	6	0,00
97	7.644.585	1	0,00

1985

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
1	22.829.087	21	0,00
2	22.829.087	18	0,00
3	22.829.087	13	0,00
4	22.829.087	2	0,00
15	22.829.087	1	0,00
16	22.829.087	9	0,00
17	22.829.087	71	0,00
18	22.829.087	617	0,00
19	22.829.087	111	0,00
20	22.829.087	22	0,00
21	22.829.087	440	0,00
22	22.829.087	5	0,00
23	22.829.087	8	0,00
24	22.829.087	23	0,00
25	22.829.087	12,470	0,05
29	22.829.087	4	0,00
30	22.829.087	1	0,00
32	22.790.749	2	0,00
33	22.774.968	1	0,00
34	22.753.498	1	0,00
35	22.724.150	20	0,00
36	22.687.141	552	0,00
37	22.670.791	20	0,00
38	22.648.320	23	0,00
39	22.618.400	57	0,00
40	22.580.056	26	0,00
41	22.531.209	22	0,00
42	22.509.606	8	0,00
43	22.480.866	14	0,00
44	22.441.602	1	0,00
46	22.314.959	2	0,00
47	22.283.958	5	0,00
48	22.242.194	101	0,00
49	22.184.150	1	0,00
50	22.106.691	2	0,00

se nadaljuje

nadaljevanje priloge I - 1985

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
51	22.000.371	3	0,00
52	21.957.210	4	0,00
53	21.899.061	7	0,00
54	21.818.989	10	0,00
55	21.709.241	66	0,00
56	21.558.565	15,600	0,07
57	21.499.680	180	0,00
58	21.419.122	1,372	0,01
59	21.314.550	118	0,00
64	20.667.832	422	0,00
65	20.500.900	12	0,00
66	20.282.040	21	0,00
67	20.194.500	16	0,00
68	20.078.723	81	0,00
69	19.919.765	29	0,00
70	19.714.923	36	0,00
71	19.441.431	1	0,00
72	19.331.363	749	0,00
73	19.183.055	915	0,00
74	18.988.529	44	0,00
76	18.422.517	1	0,00
77	18.293.294	8	0,00
78	18.125.144	1	0,00
79	17.890.505	906	0,01
80	17.605.431	2,235	0,01
81	17.240.784	3	0,00
82	17.097.935	52	0,00
84	16.667.273	1	0,00
86	15.995.960	8	0,00
87	15.831.486	10	0,00
88	15.632.434	5	0,00
89	15.379.321	89	0,00
90	15.070.468	3	0,00
91	14.686.083	21	0,00
92	14.526.580	10	0,00

se nadaljuje

nadaljevanje priloge I - 1985

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
97	13.183.066	2	0,00
123	6.761.584	1	0,00

P15

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
1	10.484.187	16	0,00
2	10.484.187	15	0,00
3	10.484.187	15	0,00
6	10.484.187	21	0,00
17	10.484.187	31	0,00
18	10.484.187	294	0,00
19	10.484.187	36	0,00
20	10.484.187	25	0,00
21	10.484.187	68	0,00
22	10.484.187	2	0,00
24	10.484.187	31	0,00
25	10.484.187	31	0,00
26	10.484.187	10	0,00
35	10.437.036	3	0,00
36	10.420.136	552	0,01
37	10.412.545	3	0,00
38	10.402.078	1	0,00
39	10.387.828	44	0,00
40	10.370.230	28	0,00
41	10.347.906	3	0,00
42	10.337.176	15	0,00
43	10.322.751	5	0,00
44	10.303.199	1	0,00
45	10.277.050	57	0,00
46	10.242.506	4	0,00
47	10.226.635	19	0,00
48	10.205.576	87	0,00
49	10.177.060	7	0,00
50	10.139.487	7	0,00
51	10.089.024	185	0,00
52	10.066.402	9,919	0,10
54	9.996.632	19	0,00
56	9.873.047	6	0,00
57	9.841.826	93	0,00

se nadaljuje

nadaljevanje priloge I - P15

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
58	9.800.679	618	0,01
59	9.747.465	13	0,00
65	9.331.654	12	0,00
66	9.218.686	7	0,00
67	9.170.246	17	0,00
68	9.105.644	17	0,00
72	8.712.518	817	0,01
73	8.637.264	817	0,01
74	8.538.717	10	0,00
79	7.989.266	678	0,01
80	7.849.368	750	0,01
82	7.602.325	17	0,00
86	7.068.260	36	0,00
87	6.989.967	5	0,00
88	6.894.878	4	0,00
89	6.773.865	44	0,00
91	6.444.002	2	0,00
92	6.371.009	96	0,00
93	6.277.114	12,813	0,20
97	5.774.681	1	0,00
102	5.201.001	1	0,00
105	4.874.180	1	0,00
106	4.709.627	3	0,00
110	4.324.260	10	0,00
111	4.155.064	77	0,00
120	3.297.670	4	0,00
121	3.146.984	1	0,00
123	3.024.888	10	0,00

P55

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
1	9.578.529	6	0,00
2	9.578.529	6	0,00
3	9.578.529	6	0,00
6	9.578.529	17	0,00
15	9.578.529	1	0,00
17	9.578.529	24	0,00
18	9.578.529	295	0,00
19	9.578.529	48	0,00
20	9.578.529	37	0,00
21	9.578.529	39	0,00
22	9.578.529	6	0,00
24	9.578.529	24	0,00
25	9.578.529	24	0,00
26	9.578.529	7	0,00
35	9.541.269	2	0,00
36	9.527.770	513	0,01
38	9.514.050	3	0,00
39	9.503.238	42	0,00
40	9.489.331	30	0,00
41	9.471.547	3	0,00
42	9.463.408	5	0,00
43	9.452.482	4	0,00
45	9.417.047	33	0,00
46	9.389.222	5	0,00
47	9.377.396	26	0,00
48	9.361.437	78	0,00
49	9.339.095	1	0,00
50	9.309.663	8	0,00
51	9.268.614	176	0,00
52	9.251.365	9,067	0,10
54	9.197.075	27	0,00
55	9.154.788	2	0,00
56	9.097.649	9	0,00
57	9.073.476	90	0,00
58	9.041.159	610	0,01

se nadaljuje

nadaljevanje priloge I - P55

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
59	8.998.323	8	0,00
65	8.659.186	23	0,00
66	8.564.123	13	0,00
67	8.525.008	22	0,00
68	8.472.556	28	0,00
69	8.401.456	1	0,00
70	8.310.201	3	0,00
72	8.142.172	778	0,01
73	8.079.085	777	0,01
74	7.995.677	7	0,00
76	7.751.358	1	0,00
77	7.694.508	1	0,00
79	7.521.635	679	0,01
80	7.399.290	707	0,01
81	7.242.933	1	0,00
82	7.179.269	12	0,00
83	7.095.689	1	0,00
86	6.695.846	41	0,00
87	6.624.969	14	0,00
88	6.538.800	11	0,00
89	6.427.891	47	0,00
90	6.289.583	1	0,00
91	6.119.549	1	0,00
92	6.051.267	70	0,00
93	5.963.546	12,153	0,20
97	5.486.363	1	0,00
101	4.998.772	1	0,00
105	4.615.879	2	0,00
110	4.082.661	11	0,00
111	3.918.444	65	0,00
120	3.107.569	1	0,00
122	2.915.781	1	0,00
123	2.854.610	1	0,00

REC

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
1	21.925.931	5	0,00
2	21.925.931	5	0,00
3	21.925.931	5	0,00
6	21.925.931	3	0,00
17	21.925.931	5	0,00
18	21.925.931	64	0,00
19	21.925.931	8	0,00
20	21.925.931	4	0,00
21	21.925.931	10	0,00
22	21.925.931	1	0,00
24	21.925.931	5	0,00
25	21.925.931	5	0,00
26	21.925.931	2	0,00
35	21.831.794	3	0,00
36	21.809.822	115	0,00
39	21.739.989	11	0,00
40	21.716.954	4	0,00
42	21.669.690	1	0,00
45	21.589.502	6	0,00
47	21.526.262	7	0,00
48	21.494.631	13	0,00
49	21.459.184	1	0,00
51	21.370.398	49	0,00
52	21.334.410	2,093	0,01
54	21.201.057	5	0,00
56	20.312.817	1	0,00
57	20.075.189	21	0,00
58	19.921.694	117	0,00
59	19.807.927	2	0,00
65	19.342.765	3	0,00
67	19.209.250	2	0,00
68	19.147.765	3	0,00
72	18.873.947	153	0,00
73	18.811.481	153	0,00
74	18.742.772	2	0,00

se nadaljuje

nadaljevanje priloge I - REC

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
79	18.385.786	135	0,00
80	18.303.643	142	0,00
82	18.136.941	3	0,00
86	17.749.133	8	0,00
88	17.542.592	1	0,00
89	17.422.212	12	0,00
91	17.132.477	1	0,00
92	16.982.623	12	0,00
93	16.809.088	2,672	0,02
110	871.810	3	0,00
111	837.649	11	0,00
123	604.103	2	0,00

T2

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
1	99.049.599	1	0,00
2	99.049.599	1	0,00
4	99.049.599	5	0,00
5	99.049.599	15	0,00
6	99.049.599	260	0,00
7	99.049.599	18	0,00
8	99.049.599	3	0,00
9	99.049.599	2	0,00
12	99.049.599	28	0,00
13	99.049.599	131	0,00
14	99.049.597	129	0,00
15	99.049.590	410	0,00
16	99.013.795	134	0,00
17	98.995.143	16	0,00
18	98.972.548	15	0,00
19	98.943.432	85	0,00
20	98.904.281	199	0,00
21	98.847.555	13,131	0,01
22	98.823.661	12	0,00
23	98.792.620	1,738	0,00
24	98.750.687	197	0,00
25	98.694.213	193	0,00
26	98.613.932	72	0,00
27	98.578.858	174	0,00
28	98.535.327	18	0,00
29	98.478.104	19	0,00
30	98.404.050	2	0,00
32	98.256.793	28	0,00
33	98.197.037	6	0,00
34	98.111.590	253	0,00
35	98.004.146	436	0,00
36	97.864.954	2,644	0,00
37	97.802.232	102	0,00
38	97.720.638	42	0,00
39	97.611.505	613	0,00

se nadaljuje

nadaljevanje priloge I - T2

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
40	97.477.514	207	0,00
41	97.307.113	103	0,00
42	97.230.496	23	0,00
43	97.128.754	18	0,00
44	96.991.840	80	0,00
45	96.810.229	12	0,00
46	96.566.594	19	0,00
47	96.457.054	156	0,00
48	96.318.931	1,388	0,00
49	96.135.300	4	0,00
50	95.890.572	5	0,00
51	95.560.148	32	0,00
52	95.420.512	46	0,00
53	95.236.228	7	0,00
54	94.991.576	169	0,00
55	94.666.201	125	0,00
56	94.218.184	3,062	0,00
57	94.023.981	2,238	0,00
58	93.766.463	1,475	0,00
59	93.432.644	208	0,00
60	92.989.937	15	0,00
61	92.405.163	30	0,00
62	92.143.841	27	0,00
63	91.804.946	672	0,00
64	91.370.793	1,040	0,00
65	90.831.861	90	0,00
66	90.137.155	255	0,00
67	89.821.557	119	0,00
68	89.430.711	226	0,00
69	88.918.399	366	0,00
70	88.267.377	502	0,00
71	87.440.233	430	0,00
72	87.059.789	4,219	0,00
73	86.588.090	4,202	0,00
74	85.986.016	50	0,00

se nadaljuje

nadaljevanje priloge I - T2

Pozicija	Pokritost	Število dvoumnih baz	% dvoumnih baz
75	85.226.524	7	0,00
76	84.260.128	93	0,00
77	83.845.653	5	0,00
78	83.307.351	19	0,00
79	82.609.806	7,199	0,01
80	81.736.203	5,043	0,01
81	80.598.231	2	0,00
82	80.070.977	681	0,00
83	79.401.195	3	0,00
84	78.573.718	33	0,00
85	77.522.286	9	0,00
86	76.201.797	270	0,00
87	75.432.401	283	0,00
88	74.546.402	23	0,00
89	73.409.126	339	0,00
91	51.619.538	8	0,00
92	51.171.530	30	0,00
93	50.601.290	41,221	0,08
97	47.285.683	4	0,00
103	42.575.686	1	0,00
104	41.751.225	5	0,00
107	38.867.406	22	0,00
108	38.191.011	13	0,00
109	37.317.506	27	0,00
110	36.232.765	6	0,00
111	34.764.611	39	0,00
116	30.296.505	1	0,00
117	29.822.694	5	0,00
119	28.458.130	18	0,00
120	27.478.474	1	0,00
121	26.213.390	13	0,00
122	25.776.625	46	0,00
123	25.208.480	258	0,00
124	24.473.081	6	0,00
125	23.537.413	109	0,00