Python for Data Analysis

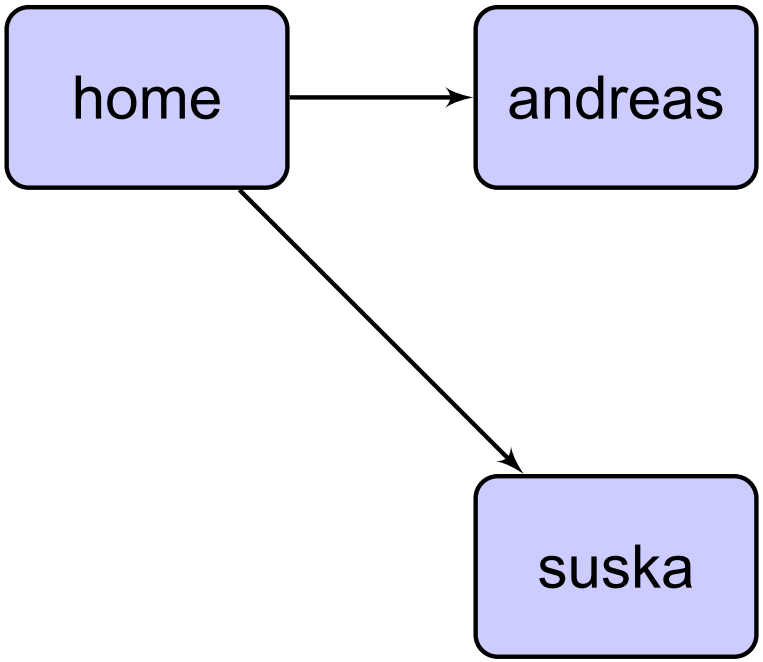
Andreas Weller, PhD

WTCHG - NHS

23.4.2014

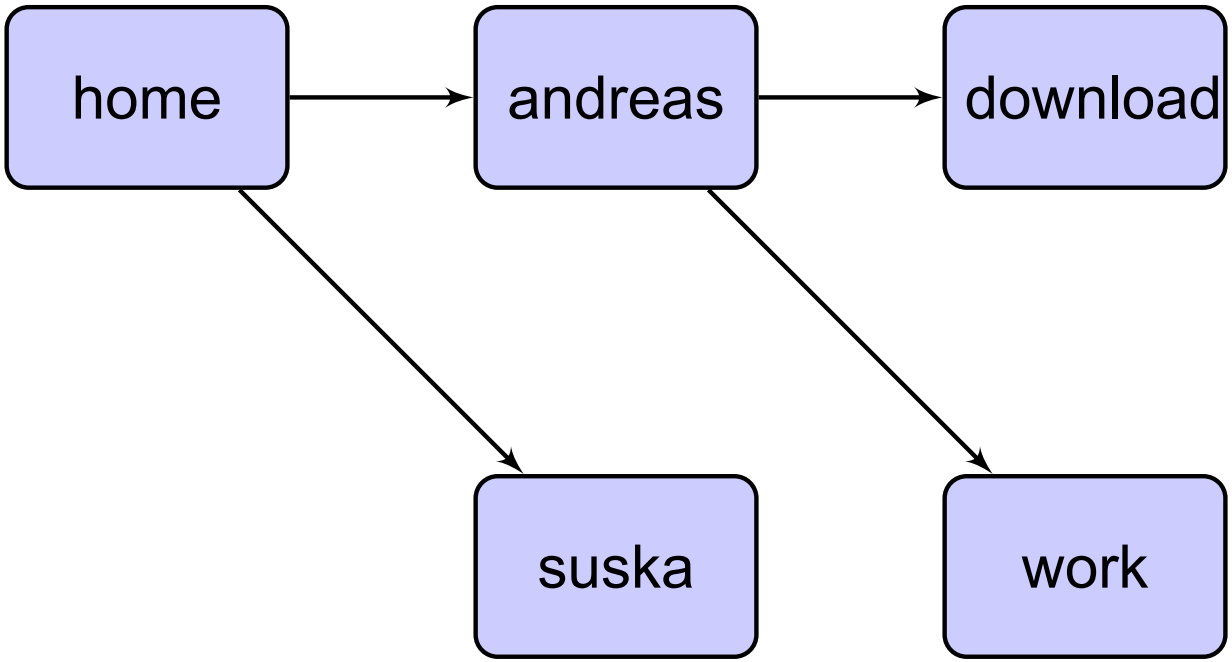
UNIX commandline

/home/andreas/work/



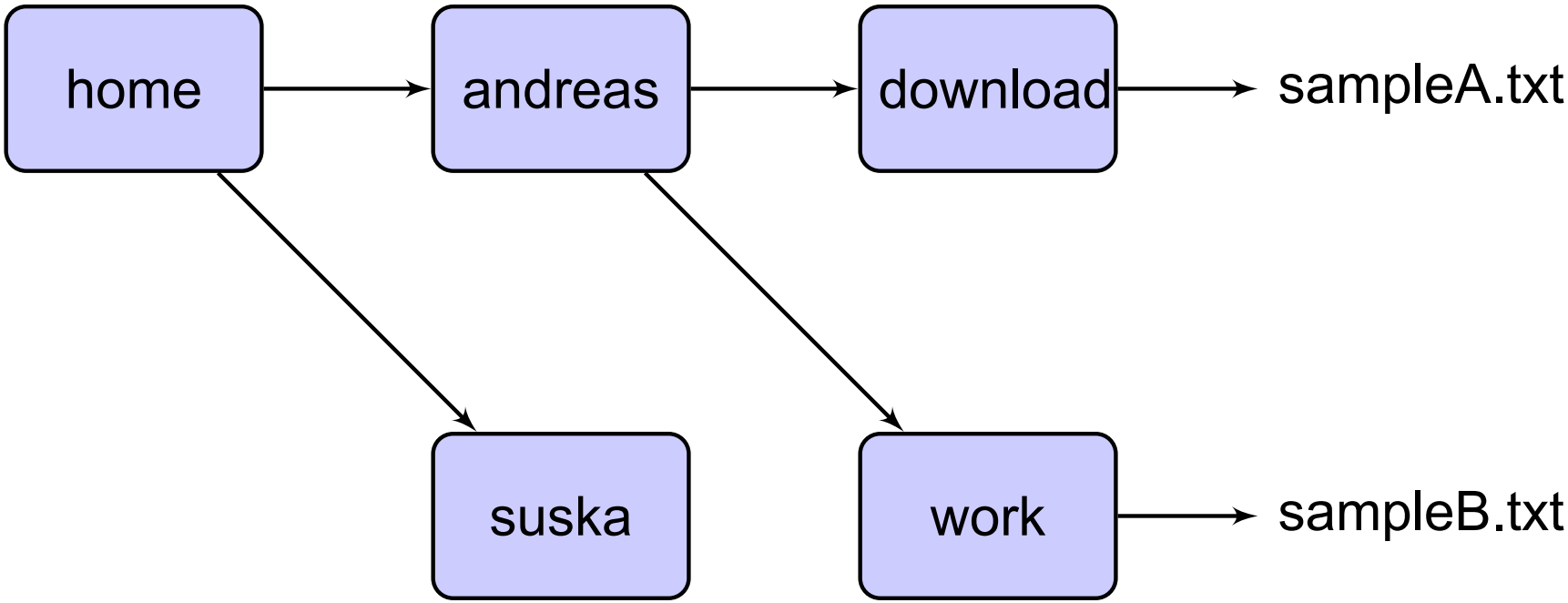
/home/andreas/downloads/sampleA.txt
/home/andreas/work/sampleB.txt

/home/andreas/work/



/home/andreas/downloads/sampleA.txt
/home/andreas/work/sampleB.txt

/home/andreas/work/



/home/andreas/downloads/sampleA.txt
/home/andreas/work/sampleB.txt

Path abbreviations

/	root
~	home
./	curr. dir
../	above curr. dir

Path names

Absolute path names

- Start from root: /
- Independent of current position
- E.g /home/andreas/work/sampleA.txt

Path names

Absolute path names

- Start from root: /
- Independent of current position
- E.g /home/andreas/work/sampleA.txt

Relative path names

- Start from current dir: ./
- Dependent of current position
- E.g ./work/sampleA.txt

Shortcuts

Shortcut	Abbreviation
man cd	show manual on 'cd'
[up]	show last command
[tab]	auto-complete
[Ctrl] + R term	search history for term
[Ctrl] + C	stop current process

Movement

Command	Abbreviation	Meaning
cd X	change dir	move to X
pwd	print working dir	show location
ls	list	show contents

File manipulation

Command	Abbreviation	Meaning
mkdir X	make dir	create dir X
touch X	-	create file X
*	wildcard	matches any character
rm -r X	remove	remove X
mv X Y	move	move X to Y

Investigate files I

Command	Abbreviation	Meaning
ls -hl	list -human -long	show file info
wc -l	wordcount -line	count lines
cat X Y	concatenate	print X and Y
X | Y	pipe	redirect X into Y

Investigate files II

Command	Abbreviation	Meaning
head	-	show first rows
tail	-	show last rows
less	-	show slowly
nano	-	open in editor

Extract and filter

Command	Abbreviation	Meaning
grep term	-	show rows with term
grep -v term	-	show rows without term
cut -f 2	cut -field	show 2nd column
cut -c 1-10	cut -character	show first 10 characters

Extract & count unique entries

Command	Abbreviation	Meaning
sort	-	sort rows alphabetically
sort -n	sort -number	sort rows numerically
sort -k2,3n	sort -column -number	sort on 2nd and 3rd column
uniq	-	only show unique entries
uniq -c	uniq -count	count unique entries

Manipulate and filter

Command	Meaning
sed 's/old/new/g'	change 'old' to 'new'
rename 's/old/new/g' *.vcf	change filenames
awk '$1 < 5'	filter on value in 1st column
awk 'length($5) == 1'	filter on length of 5th column
awk 'print $0"\extra"'	append a column "extra"

Common pipes

Count chromosomes in vcf

```
cut -f 2 file.vcf | sort | uniq | wc -l
```

Common pipes

Count chromosomes in vcf

```
cut -f 2 file.vcf | sort | uniq | wc -l
```

Find most common variant type

```
cut -f 4,5 file.vcf | sort | uniq -c | sort -n | tail
```

Common pipes

Count chromosomes in vcf

```
cut -f 2 file.vcf | sort | uniq | wc -l
```

Find most common variant type

```
cut -f 4,5 file.vcf | sort | uniq -c | sort -n | tail
```

Extract exon/UTR variants in TP53

```
grep TP53 file.vcf | grep -v intron | grep -v intergenic > output.vcf
```

Common pipes

Count chromosomes in vcf

```
cut -f 2 file.vcf | sort | uniq | wc -l
```

Find most common variant type

```
cut -f 4,5 file.vcf | sort | uniq -c | sort -n | tail
```

Extract exon/UTR variants in TP53

```
grep TP53 file.vcf | grep -v intron | grep -v intergenic > output.vcf
```

Extract high quality Indels

```
awk 'length($4) > 1' file.vcf | awk '$6 > 40' > output.vcf
```