**Contiuous protocol EPPN project data analysis**

* ASM: our codes were a bit messy for the allohexaploids! It starts to get rather complex at the point we have six parents, which I suppose is even more reason for us to have a good naming system on our side :p. I used the PH codes to describe the comparisons anyway on the grounds this is likely the easiest for the analysis.
* parent PH04 and PH05 were the 'cauliflowers' that failed to flower, so there will be rosette and digital biomasss for them but no pod or seed yield data. All other genotypes should have full data sets. (Fiona)
* We should be able to supply data for several of these without much development work; from my perspective it would be useful to have someone do a sense check on the initial data (general morphology) to make sure we're not making any fundamental errors prior to committing to deeper analysis. (Kevin)

**Meeting with Annaliese, 21-1-22**

1. Are all parents homozygous, if not otherwise stated? Yes! Varieties like Boomer are homozygous. Exception: B578 is heterozygous. Allohexaploids are also heterozygous, except for hexaploids produced from a tetraploid and a diploid parent (Mwathi, 2020)
2. We want to separate effects of ploidy level and heterozygosity on growth and fertility.
   1. Do we also have to separate the effects of ploidy level and the number of different genomes? E.g. tetraploids with 2 or 3 genomes. => probably relevant for the proposed advantages of polyploidy (gene redundancy, heterosis, cell size). No, we cannot clearly separate effects of ploidy level and heterozygosity. Think of this as comparing polyploids with their di-/tetraploid parents. See set of comparisons. In BBAC tetraploids the A and C genome behave as one set (see figure in pptx). BBAC can be thought of as having two genomes / two sets of chromosomes. Wenn es in der Genexpression zwischen A und C Probleme gibt, sieht man das an einem stark verringerten Wachstum dieses Individuums.
   2. Why are there no autopolyploids included? Autopolyploids can be produced from one or two parents, right? (=different degrees of heterozygosity) No seeds. Recently created. Unstable.
3. Attribution of traits to genomes: Is it possible to attribute the incorporated traits (growth and reproductive properties) to the A, B and C genome of the diploid or tetraploid parents? Attribution theoretically possible for qualitative traits (like resistance in B. nigra or white flower coloure in B. oleracea) but not for quantitative traits such as fertility traits (number of siliques and seeds).
4. Do I have to know in more detail, how our genotypes were made (original papers)? No! See figure in presentation. Evtl. Mwathi 2020 lesen.
5. Genomic composition of B. juncea x B. carinata tetraploids: BBAC or different combinations? Different combinations (see Katche, 2019, New Phytol.)
6. comparison 4: the 7 allohexaploid lines from *B. napus* x *B. carinata* x *B. juncea* combinations => Are they produced from the natural or the synthetic *B. napus*? I.e. do we have a pyramid in this case? No, they are produced from natural *B. napus.*

How are parents made homozygous? Inbreeding or microspore culture, which produces double haploid lines. Which is the common one in our work and which in commercial plant breeding?

7. Growth/yield measurements: Are the resulting tetraploids and hexaploids all oil crops? So, none of them are root or head forming vegetables? => Hybrid vigour in terms of larger organs. None of them are root or head forming vegetables, as far as she knows. Some napus are canola, i.e. oil crops. Carinata and juncea are condiment crops. Oleracea is weedy.

8. `Average of branch #´ = branch number/plant or branch length? Probably number. Check for trade-off between seed formation and growth?

9. `Average pod number´ = per plant? Yes! Probably per plant

10. Do the barcodes refer to individual plants? Ask collaborators in meeting

11. Are the fertility parameters assessed here directly related with the stability of meiosis in segregating progeny? No direct relationship, but meiotic instability can cause reduced number of seeds and chromosome losses can lead to growth retardation. Keep bias in mind: Fertility as assessed here as a parameter/indicator of hybrid vigour, but fertility is also affected by meiotic instability in newly formed hybrids.

12. Has the growth rate also been measured manually or by imaging? Probably both, because they wanted to train the imaging system.

13. Randomized block design: How many blocks? Figure of the design? Location 1 and 2? Comp 5 and 6? Ask collaborators in meeting.

14. `Growth rate´, `Total biomass´ in terms of weight and `Time to flowering´ not found in table => assessed by automated phenotyping plattform? Ask collaborators in meeting

15. Genomic and genotyping data mentioned in proposal supposed to be used for further analyses? No specific use, because genes responsible for growth, yield and fertility (=hybrid vigour) of hybrids unknown.

**Meeting with Annaliese, 22-02-22**

* What is this paper supposed to cover? Research questions from your excel sheets:
  + parent/progeny comparisons,
  + species comparisons and
  + second level analyses for similarity of trends.
* What are the most relevant growth and fertility parameters from this data set (manual measurements)?
  + Fiona will provide seeds/10 pods and total seeds/plant data soon
  + Relevant growth and fertility parameters => check paper (Mason: Agricultural selection and presence–absence variation in spring-type canola germplasm => effects of polyploidy and hybridization on these traits?)
  + To get all the parameters proposed in the proposal, imaging data have to be integrated
  + Henrik will send list of agronomically important traits
  + Best parameters:
    - Height, biomass, growth rate, seeds/pod, seeds/plant, days to flowering
    - Best fertility parameter: seeds/pod. But no heterosis effect on this trait expected.
    - Plant height/biomass: heterosis effect expected
  + Check Danis photos to know differences in phenotypes, particularly of the hexaploids
* Done up to now: Boxplots, descriptive statistics, ANOVAs and ANOVAs for randomized block design for comparisons 1 and 2 (A-Y). How to deal with non-normal distributions?
  + No worries
  + But if we are talking about significance, we need to perform a Kruskal-Wallis-Test
    - Der Kruskal-Wallis-Test (nach William Kruskal und Wilson Allen Wallis; auch H-Test) ist ein parameterfreier statistischer Test, mit dem im Rahmen einer Varianzanalyse getestet wird, ob unabhängige Stichproben (Gruppen oder Messreihen) hinsichtlich einer ordinalskalierten Variable einer gemeinsamen Population entstammen.
  + Some traits are not normally distributed => google
    - Examples
      * Seed set (many sterile plants)
      * Flower colour
    - Quantitative genetic effect
* Next steps: Similar tends between comparison groups for specific traits (second level analysis)?
  + Heat map: Difference parent mean vs. progeny => transgressive segregation? (most interesting phenomenon)
  + Less important: maternal effects = heat map with maternal genotypes as references => differences mother vs progeny
    - Mother = first genotype in table? Check Danis list (email)
  + Check correlation between traits
  + Do PCA for traits
* What are we going to show in the paper: first or second level analyses?
  + Mostly second level analyses
  + First level only examples, eg 5 out of 20 traits show transgressive segregation (write in text, no figure)
* Developmental data: day at which BBCH51 (flower buds visible from above) and BBCH61 (10% of flowers on the main raceme open) have been reached?
* Any joint analysis with imaging data planned or is that going to be a separate paper?
  + Aberystwith prefers to get all the data together first and then decide about den publications
  + We are possibly going to integrate field data from Bonn => discuss with Henrik

**Meeting 01-03-22**

**Mail ASM 7.3.22**

Hi Helen,

my recommendation on traits is to **first identify which traits show which trend for which comparisons**, and summarise all this information in text form (like a results paragraph). **First level is the comparisons listed**, but **second level is really within the comparison groups**, and it may not make sense to combine absolutely everything together in one table or figure (across groups would maybe be a third-level comparison). This will depend on the results from the individual and within-group comparisons, but we may not have similar trends. Once this is all sorted, we can pick which traits it makes sense to show as examples in figure etc., but all the data has to be analysed first so we know what is happening with each trait and genotype combination. So a written summary as it would show up in the Results section of the paper may be the most effective now, alternatively a large supplementary table showing everything at first might be helpful if this is at all feasible.

Best wishes,

Annaliese

Hi Helen,

looks like a nice format! Two suggestions: firstly we need to just show the **direction of the effect overall between the parents and synthetic/s, not all plants together** (so one line = one comparison, not one genotype in the heatmap). For this I would also recommend to first check all comparisons within a comparison group, so all the synthetic napus first, then all the allohexaploids separately, etc., then summarise all these results before combining them all together.

Secondly I'm afraid we really need to switch to the understandable codes from the PH codes. We need to use the PH codes to communicate with Kevin and the Aberystwyth group because all of their data is already collected using these codes, but for "final" figures or figures for discussion within our group we need to use the genotype codes. These are just impossible to read otherwise, I don't think it's sensible to keep them for the figures.

Attached in purple the recommended codes! Hope that's not too annoying to change, but then at least we can see the combinations on the figures...

Thanks!

Annaliese

**Joint analysis of manually assessed and imaging data**

* Comparison groups
  + Comparison group 1 (only A available): resynthesised *Brassica napus*
    - Parents: *B. rapa*, *B. oleracea*
    - Progeny: resynthesised *B. napus*
  + Comparison group 2: novel allohexaploids
    - Parents: *B. napus* x *B. carinata* x *B. juncea*
    - Progeny: novel NCJ allohexaploids
  + Comparison group 3: JC hybrids
    - Parents: *B. juncea* x *B. carinata*
    - Progeny: JC hybrids
      * Different genomic compositions (see Katche, 2019, New Phytol.) – relevant?
  + Comparison group 4: Allohexaploid hybrids
    - Parents: NCJ allohexaploids x JO allohexaploids
    - Progeny: F1 allohexaploids

**Sets of comparisons: ploidy level, number of genomes and heterozygosity**

comparison 1: *B. rapa*, *B. oleracea*, synthetic *B. napus* from known parent genotype combinations

* AA (diplod, homozygous) x CC (diploid, homozygous) = AACC (tetraploid, 2 genomes, homozygous)
* Effects of ploidy level and number of genomes (=allopolyploids)

comparison 2: *B. juncea* J1, *B. carinata* C1 and C2, and the *B. juncea* x *B. carinata* tetraploids

* AABB (tetraploid, 2 genomes, homozygous**?**) x BBCC (tetraploid, 2 genomes, homozygous**?**) = BBAC**?** (tetraploid, **3** genomes, homozygous)
* Effects of number of genomes. Same ploidy level.

comparison 3: *B. oleracea* TO1000, *B. juncea* J3, and their allohexaploid O1J3

* CC (diploid, homozygous) x AABB (tetraploid, 2 genomes, homozygous**?**) = AABBCC (hexaploid, 3 genomes, homozygous**?**)
* Effects of ploidy level (= number of different genomes, in this case)

comparison 4: the 7 allohexaploid lines from *B. napus* x *B. carinata* x *B. juncea* combinations, with their *B. napus*, *B. juncea* and *B. carinata* parents

* Order of crossing?
* AACC x BBCC x AABB = AABBCC
* Effects of ploidy level (= number of different genomes, in this case)

a complex set of comparisons: each of the heterozygous hexaploids to their two parent hexaploid lines (and maybe all of them to the species parents, but let's see first how reasonable this is)

* AABBCC (homozygous) x AABBCC (homozygous) = AABBCC (heterozygous)
* Effects of heterozygosity. Same ploidy level / number of different genomes.
* **Questions:**
  + **First level analyses**
    - Parent-progeny comparison (all groups): Are progeny phenotypes
      * intermediate between parent phenotypes,
      * more similar to Parent 1,
      * more similar to Parent 2, or
      * do they show transgressive segregation (above or below both parent values) for each trait?
    - Within / across species/genomic composition comparisons:
      * Are resynthesised *B. napus* traits more similar to natural *B. napus* or to the parent *B. rapa* and *B. oleracea* parent lines? (Comparison group 1)
      * Do NCJ allohexaploids have more similarity to each other than to their parent genotypes? (Comparison group 2)
  + **Second level analyses**
    - Parent-progeny comparison (all groups, second level): Are there any trends between the independent comparison groups?
    - Comparisons 2 and 4 together: Do allohexaploid progeny show hybrid vigour compared to their allohexaploid parents? Do the allohexaploid parents show hybrid vigour in comparison with their diploid and tetraploid parental lines?
    - Correlation analyses: Trade-off between growth and fertility?
  + **Separate the effects of ploidy level, heterozygosity (and the number of different genomes?):** No, we cannot clearly separate effects of ploidy level and heterozygosity. Think of this as comparing polyploids with their di-/tetraploid parents.
* Faktor Zeit: Wie aufnehmen?

**Wichtigste Traits aus Annalieses Paper**

* Yield traits assessed were:
  + **number of days to 50% flowering** as estimated from seeding to initiation of flowering in ~50% of the plants in each plot (DF50);
  + **number of days from seeding to maturity** as evident from physiological maturity of about >90% pods/plot (DM);
  + **plant height** was measured as distance from ground level to the tip of main axis at cessation of flowering in cm (PH);
  + **main shoot length** measured as distance from the last primary branch to the tip of main shoot at cessation of flowering in cm (MSL);
  + **number** of productive silique bearing **primary branches** on the stem (PB);
  + **number** of productive silique bearing **secondary branches** borne on the primary branches (SPB);
  + **number of siliquae** on the main shoot (MSP);
  + average **silique length** of five silique from the middle of the main shoot in cm (SL);
  + average **seed number in 10 random siliques** (SSL);
  + **seed weight of 1000 seeds** in grams (SS);
  + **seed yield in kg/ha** (Yield);
  + foliar **chlorophyll content** measured using SPAD-502 (PP);
  + **leaf area** in cm2 (LA);
  + **leaf area index** (LAI) and
  + **percentage seed oil** (Oil) and
  + **protein content** (Protein) estimated using a Foss NIRS system

**Fragen**

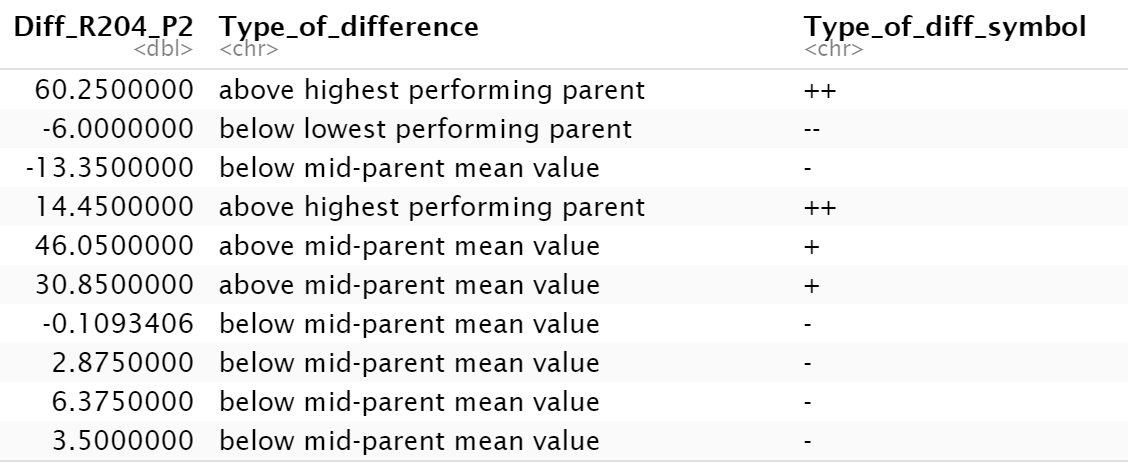
* Do I have to know and take into account the differences between the progeny sets in comparison group 2?

**Fragen an Kevin**

* What do you mean by peak? Date of the largest number of open flowers? DeepL gives funny translations, that don’t make any sense.

**Feedback from Annaliese, 14-03-22**

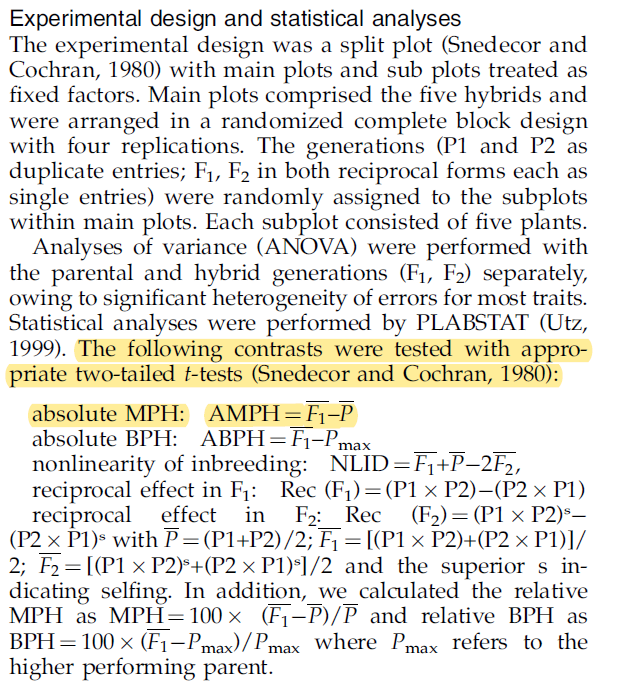
* Mein Stand zu dem Zeitpunkt:

****

* yeah at first analyse individual progeny sets, because there might be strong genotype-specific effects which we won't see otherwise
* I think I would really like to know if **above or below midparent mean value is significantly above or below midparent mean value**. But if this is a separate analysis I would cluster all as "between parents" together
* okay, then for sure the three categories I would recommend would be "hybrids between parent values", "hybrids above highest parent mean values" and "hybrids below lowest parent mean values" just to get an idea
* At the moment I'm trying to produce the summary table or text, you suggested last week. That means I've done all the ANOVAs, I'm just summarizing it all in one table. But I will check if this can be done from the ANOVA output. I just didn't know how to feed all the results into one table. I'll try again.
* I guess my expectation would be that most of the traits fall between the parent values and are not significantly different to the midparent mean. This is the baseline assumption, the H0 hypothesis
* Ok, I had done the ANOVA only for differences between individual parents and progeny sets. If I compare the parent mean to the progeny, I would expect that the joint parent's data don't match the assumptions of normality and equality of variances anymore. So, I have to use a Kruskal-Wallis Test, right? Or can I do an ANOVA anyway? Maybe I have some misconception here ...
* From what I understand ANOVA is very robust to minor violations of normality and variance distribution assumptions, I don't think there is any real expectation that this would be invalid here
* There are a couple of different ways to compare hybrids and parents that are all probably fine also
* Can you recommend a good ressource on this (comparing progeny to parent means)? Because if the distribution of the parents mean is +/- bimodal, I would prefer a different way of analyzing the data.
* I would recommend that you just use the method you were using before and ignore my comment about midparent means 😊
* but it might be a good idea to check how this is usually done in the literature for sure
* so yes I fully agree that we should avoid comparing mid-parent means to hybrids statistically using ANOVA then. It looks like there is some other way to do this via regression analysis...? But maybe stick with your original plan for now!
* Wenn ich direct Signifikanzen einfüge, die Kategorien Progeny unterschiedlich von P1, P2 und MPV bilden
* Wenn ohne Signifikanzen reicht P1, P2 und zwischen Eltern

**1. Question:** which significance test do you usually use for comparing phenotypic trait means of progeny to mid-parent values, if the latter has a bimodal distribution? That means a non-normal distribution for the joint parents data.

* Agim: if the samples size is big enough t-test or ANOVA. otherwise Kruskal-Wallis-Test
* Barth (2003): T test



Barth (2003): Heterosis for biomass yield and related traits in five hybrids of Arabidopsis thaliana L. Heynh

**2. Question: Term “negative heterosis”?**

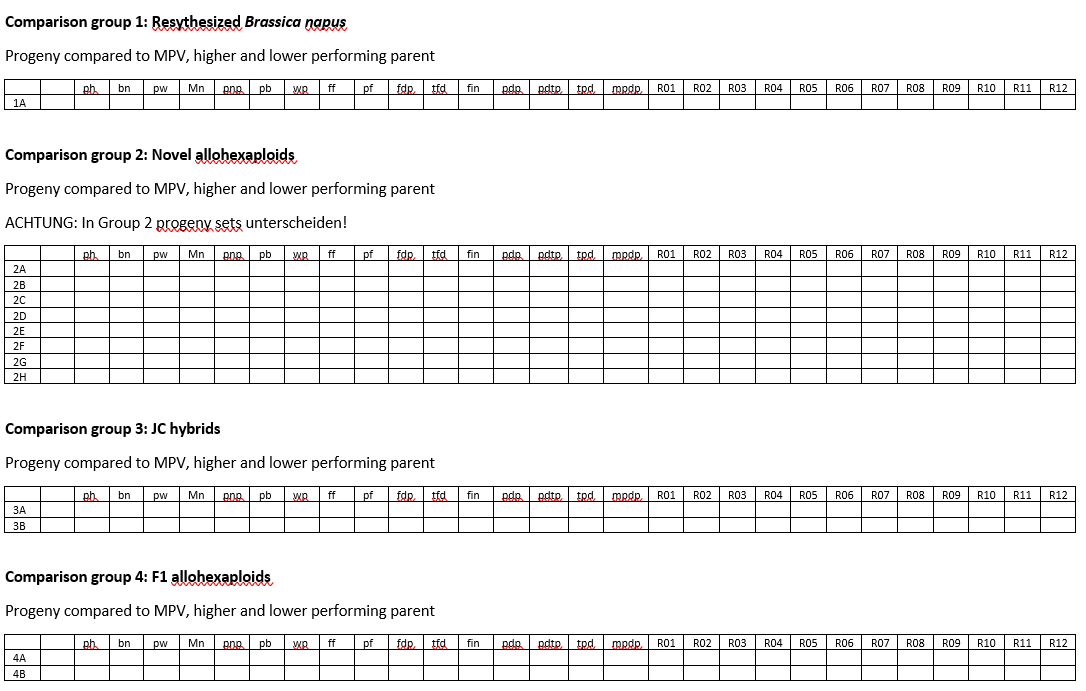
Yes, some authors use the term negative heterosis: "In this article terms ‘‘positive’’ and ‘‘negative’’ heterosis are used, meaning that the trait value is increased/decreased compared to the mean of the parents or to the better parent."

**3. Question: How to summarize outputs of statistical tests in summary table?**

Rstatix package

**Create summary table first level analyses (parent-progeny comparisons)**

**Automate this process based on output of statistical tests, 17.03.22**

1. Wie soll das Ergebnis am Ende aussehen?
2. Wie komme ich theoretisch dahin?
   1. Ausgangsdatensatz: PhenComplClean, nach Umbenennung der Variablen
   2. Genotypen der Vergleichsgruppen auswählen (dplyr)
   3. Mutter / Vater unterscheiden
   4. Genotypen innerhalb der Gruppen als P1, P2, P3 und Progeny 1, 2, 3 bezeichnen
   5. Assummptions testen
      1. Normality
      2. Equality of variances
   6. Anova
   7. Post hoc test
   8. Wenn assumptions erfüllt, p-Wert von ANOVA und post hoc test in Tabelle übertragen
   9. Wenn nicht, T Test oder Kruskal-Wallis-Test durchführen
   10. Ergebnis in Tabelle übertragen
3. Wie komme ich praktisch dahin?
   1. Für einen trait und eine gruppe gelöst!
   2. Auf alle übertragen => Fei nach Möglichkeiten fragen, dies gesammelt zu verarbeiten, nicht einzeln

**Fortsetzung: Create summary table first level analyses (parent-progeny comparisons), 18-03-22**

**To do:**

1. Seed data aufnehmen
2. BBCH data aufnehmen?
3. Fei nach Möglichkeiten fragen, dies gesammelt zu verarbeiten, nicht einzeln
4. Umsetzen

Feis code:

data = read.table('data.txt')

genotype = c('Parent1','Parent1'...) #put the genotype name. if the name is same, it means replicates.

p.summary = numeric()

Tukey.result = list()

#loop is how many traits

for(i in 1:ncol(data)){

aov.test = aov(data[,i]~genotype)

p.summary[i] = summary(aov.test)[[1]][1,5]

Tukey = TukeyHSD(aov.test)

Tukey.result[[i]] = Tukey.result$genotype

}

for(i in 1:ncol(data1)) { # for-loop over columns

data1[ , i] <- data1[ , i] + 10

}

**Erstellung Gesamttabelle First level analyses, 21.3.22**

* Erst gesamten Rohddatensatz analysiert. Dann nach Comparison groups aufgetrennt.
* Code:

sink("ANOVA-Tukeys.doc") #Export R output as txt & csv File

for(i in 5:ncol(data)){

column <- names(data[i])

AVz <- summary(aov(data[,i] ~ Genotype, data = data))

Tukey <- TukeyHSD((aov(data[,i] ~ Genotype, data = data)))

print(column)

print(AVz)

print(Tukey)

}

sink()

1:ncol(data)

AVz <- rep(NA, ncol(data)) # creates table with same number of columns to fill in

sink("ANOVA-p-values.doc")

for(i in 5:ncol(data)){

column <- names(data[i])

AVz <- summary(aov(data[,i] ~ Genotype, data = data))[[1]]["Pr(>F)"]

Tukey <- TukeyHSD((aov(data[,i] ~ Genotype, data = data)))

print(column)

print(AVz)

}

sink()

**22.3.22 Berechnung weiterer Statistiken und Gruppierung der Ergebnisse**

* Was muss noch rein?
  + Parents mean, difference from parents mean
  + Shapiro-Wilk-Test auf Normalverteilung
  + Levene’s Test auf Varianzhomogenität
  + ANOVA für randomized block design
* Was will ich am Ende haben?
  + Tabelle aller traits für alle Untervergleichsgruppen, unterteilt nach Progeny sets
  + Assumptions für alle Untervergleichsgruppen (2A, B, C, …) , unterteilt nach Progeny sets
  + Mittelwerte und p-werte für Unterschiede (ANOVA und letzteres aus Tukey test)
    - Progeny set – MPV
    - Progeny set – P1
    - Progeny set – P2

**23.3.22 Versuche Loop im Loop für data sets und Spalten**

**Datensätze:** c1A, c2A\_PH29, c2A\_PH47, c2B\_PH22, c2B\_PH42, c2C\_PH23, c2C\_PH30, c2D\_PH24, c2D\_PH35, c2D\_PH44, c2D\_PH48, c2E\_PH25, c2E\_PH31, c2E\_PH36, c2E\_PH39, c2F\_PH26, c2F\_PH40, c2H, c3A, c3B, c4A, c4B, c4C, c4D, c4E, c4F, c4G, c4H, c4I

Für jeden Datensatz

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Trait 1  ph | MW P1 | MW P2 | MW P3 | ANOVA | Tukey  Diff1  Diff 2  Diff 3 | MPV | Diff MPV |  |
| Trait 2  bn |  |  |  |  |  |  |  |  |
| Trait 3  pnp |  |  |  |  |  |  |  |  |

**2 loops:**

1. Loop: nimmt immer einen neuen Datensatz aus einer Liste oder nimmt immer neue Zeilen aus einem kompletten Datensatz
2. Loop: Wendet mehrere Funktionen (aov, TukeyHSD, …) auf jede Spalte dieses Datensatzes an

**Allgemeine Struktur eines loops:**

output <- vector("double", ncol(df)) # 1. output

for (i in seq\_along(df)) { # 2. sequence

output[[i]] <- median(df[[i]]) # 3. body

}

**EPPN summary table first level results (nach R Kurs mit neue Skills), 5.4.22**

**Plan für heute:**

* Ziel:
  + Ansatz mit apply functions (R-typisch, Empfehlung von Katja) nochmal bearbeiten, bugs entfernen, zum Laufen bringen.
  + Für kompletten Datensatz oder subsets. Aber für alle statistischen Tests, die ich brauche.
  + Wenn es bis Mittwoch Abend nicht läuft, Kevin, Fei oder R community um Rat fragen.
* Vorgehen:
  + Lernmodule dazu nochmal anschauen
  + überlegen, wie es theoretisch laufen soll
  + Wissen übertragen, Codierung überlegen
  + Advanced R Buch und Cook book zur Hilfe nehmen
  + Funktionen nutzen, um mit Bugs umzugehen
  + Stück für Stück aufbauen, erstmal Teile zum Laufen bringen, dann das Ganze

**To Do 07.04.22**

1. gewünschtes Ergebnis ganz genau aufzeichnen

2. weg dahin überlegen

3. Informationsquellen suchen (siehe unten und Videos zu nested apply functions und eigenen Funktionen) und die 3 wichtigsten lesen / anschauen (ggf. Zeitplan) und Tools rausschreiben

4. Weg überarbeiten

5. wenn’s nicht klappt, aktuellen Code an fei und Kevin. Fragen.

6. notfalls alten Loop nutzen

Apply functions versus dplyr

<https://www.r-bloggers.com/2022/03/complete-tutorial-on-using-apply-functions-in-r/amp/>

<https://statsandr.com/blog/how-to-do-a-t-test-or-anova-for-many-variables-at-once-in-r-and-communicate-the-results-in-a-better-way/>

**Frage 1: Gewünschtes Ergebnis**

H0 hypothesis: Trait not different from MPV

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison groups** | **Trait 1** | | | **Trait 2** |
| **Difference from LPV** | **Difference from MPV** | **Difference from BPV** | **…** |
| **1** | Mean difference with significance level given as \*, \*\*, \*\*\* |  |  | **…** |
| **2** |  |  |  |  |
| **3** |  |  |  |  |
| **…** |  |  |  |  |

* LPV = lower parent value
* MPV = mid parent value  
  BPV = better parent value
* for all traits
* for all individual progeny sets
* Add sex of parents
* ANOVA relatively robust
* alternative: Kruskal-Wallis oder Regression analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison groups** | **Trait 1** | | | **Trait 2** |
| **Difference from LPV** | **Difference from MPV** | **Difference from BPV** | **…** |
| **1** | % difference with significance level given as \*, \*\*, \*\*\*  Siehe Li, 2021, Genome-wide analysis |  |  |  |
| **2** |  |  |  |  |
| **3** |  |  |  |  |
| **…** |  |  |  |  |

**Frage 2: Wie komme ich dahin?**

1. Datensatz in Vergleichsgruppen splitten
   1. testen, wie ich den Input für meine eigene Funktion neu definieren kann (Variablen: P1, P2, P3, Progeny, …)
   2. testen, wie ich Listen von dataframes mit lapply auswerte
   3. Testen, wie ich iterative Prozesse verschachteln kann
2. ggf. Geschlecht der Eltern rein, maternal effects
3. Für alle 30 Merkmale/Parameter
   1. Mittelwert der Eltern bilden
   2. (Differenzen von BPV, MPV und LPV bilden, macht Post hoc Test)
   3. ANOVA assumptions testen
      1. Shapiro-Wilk-Test auf Normalverteilung
      2. Levenes Test auf Varianzhomogenität
         1. testen, wie ich mit apply functions, map oder dplyr functions statistische Tests auf mehrere Traits/columns anwenden kann
   4. ANOVA für Vergleichsgruppen
   5. Post hoc tests
   6. ggf. Kruskal-Wallis-Test oder Regression analysis
4. Aus dem Output des Post hoc tests (Vector, Matrix, Liste, Dataframe?) entnehmen und ich neuen Vectoren zusammenführen:
   1. Differenz zu BPV mit Signifikanzlevel in Sternchen
   2. Differenz zu MPV mit Signifikanzlevel in Sternchen
   3. Differenz zu LPV mit Signifikanzlevel in Sternchen
      1. herausfinden, wie ich zB p-Werte aus Output entnehme
5. Die Vectoren für einzelne Traits in Gesamttabelle zusammenführen

**Frage 4: Wie programmiert man das?**

* Direkt neue Tabelle erzeugen oder einzelne Vectore und diese zusammenbinden?
* herausfinden, wie ich
  + den Input für meine eigene Funktion neu definieren kann (Variablen: P1, P2, P3, Progeny, …)
  + Listen von dataframes mit lapply auswerte
  + iterative Prozesse verschachteln kann
  + mit apply functions, map oder dplyr functions statistische Tests auf mehrere Traits/columns anwenden kann
  + z.B. Differenzen und p-Werte aus Output entnehme

**Frage 3: Quellen konsultieren**

* r bloggers (<https://www.r-bloggers.com/2022/03/complete-tutorial-on-using-apply-functions-in-r/amp/>)

# Calculating the mean for each row in the data frame

row.avg <- apply(X = example[, 2:4], MARGIN = 1, FUN = mean)

# Pivot the data so that the data are in long format instead of wide format

example <- pivot\_longer(example, cols = 2:4, names\_to = "time", values\_to = "height")

# Use tapply() to find average height by time grouping

tapply(X = example$height, INDEX = example$time, mean)

* Soetewey: How to do a t-test or ANOVA for more than one variable at once in R (<https://statsandr.com/blog/how-to-do-a-t-test-or-anova-for-many-variables-at-once-in-r-and-communicate-the-results-in-a-better-way/>)



works!

**13.04.22**

Ziel: code fertigstellen, ANOVA, assumptions und posthoc für dataframes und Spalten

1. morgens die wichtigsten Videos zu eigenen Funktionen und apply/map funktions ansehen

2. Was will ich haben? In mein Protokoll schauen! Was hat bereits funktioniert? Vielversprechendster und state of the Art Ansatz?

3. Meinen besten Versuch entwickeln und als frage an Kevin, Fei und Stackoverflow und super Data Science

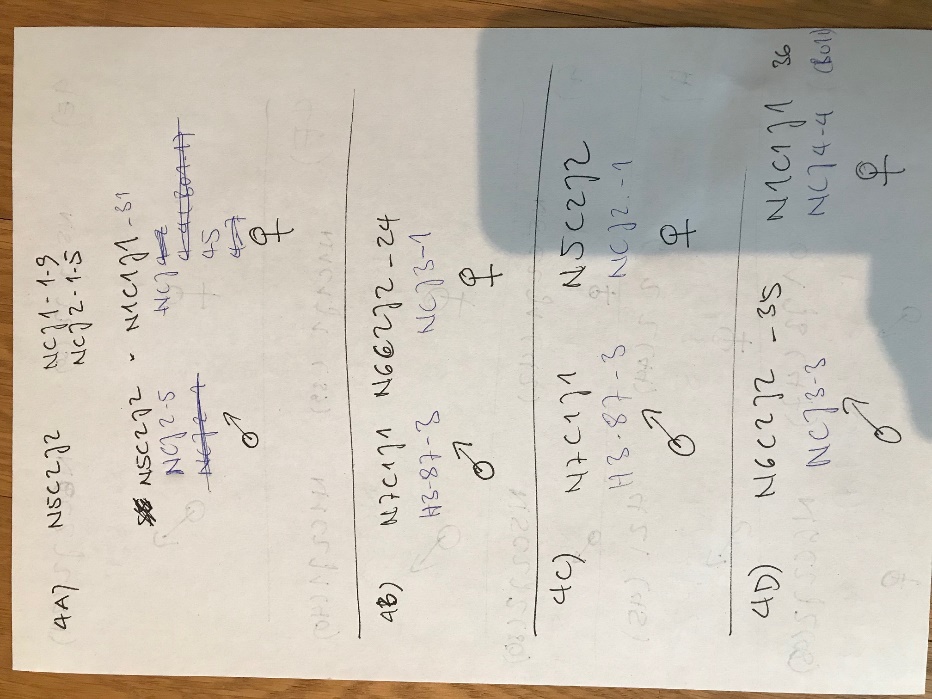
**14.4.22**

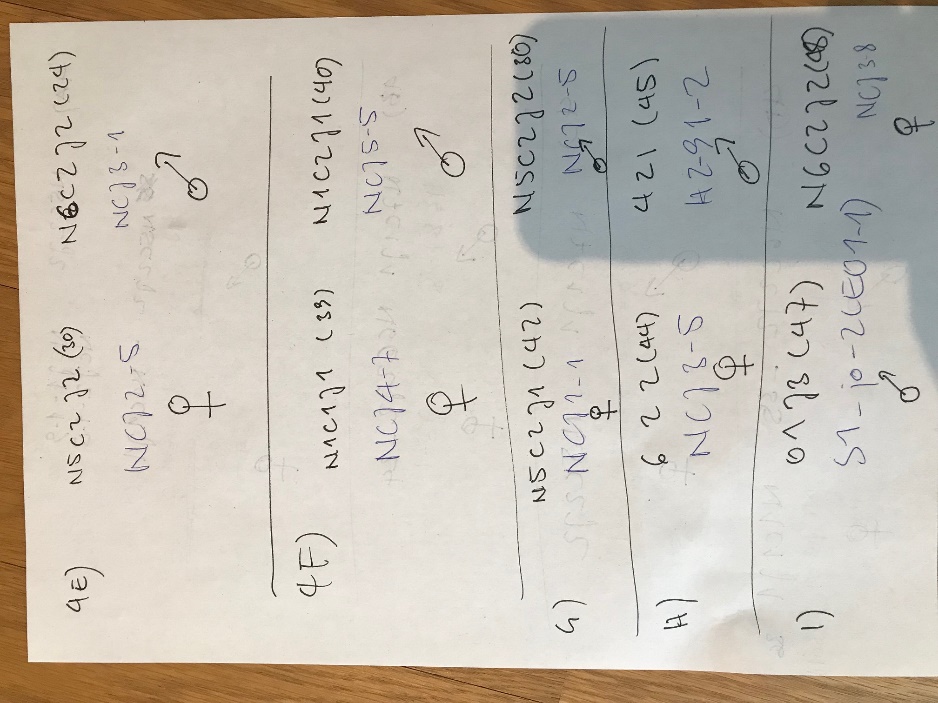
* Für P1, P2, P3 und progeny (mehrere Gruppen) ANOVA
* Für MPV und progeny (2 Gruppen) T-Test. Generell nur T-Tests machen, weil es immer um einzelne comparisons geht?
* Testergebnisse in einer Tabelle zusammenführen.
* Wenn ich funktionierende Vorlagen finde, conditionals einbauen.
* Versuchen, das Ganze für die Liste der dataframes zu machen. Dazu Videos zu Listen, apply und eigenen Funktionen.

**Geschlecht der Eltern**

Annaliese:

* **Brassica rapa** is always the mother parent for the Abel synthetic lines (=synthetic B. napus).
  + R1 and R2
* For the synthetic B. juncea x B. carinata the **B. juncea** is the mother (J1C1 and J1C2).
* And for all NCJ allohexaploids the combination is B. napus x B. carinata to make an F1 hybrid, then this hybrid is crossed with B. juncea, so in principle the **B. napus** is the mother and I would write it in the order B. napus, B. carinata, B. juncea for the parents (hence NCJ).
* For the B. oleracea x B. juncea the **B. oleracea** is the maternal parent.
* This information is all published in papers from our group and from Heiko Becker's group in Goettingen, but I'm trying to get all the information into the shared drive folder "Plant material and crossing"





**Code defining mother and father genotype**

# subsetting & defining mother / father genotypes

# P1 = mother

#1A

comp\_1A <- DATA[DATA$Genotype %in% c("R2", "O4", "R204"), ]

comp\_1A <- comp\_1A %>%

mutate(Generation = case\_when(

Genotype == "R2" ~ "P1",

Genotype == "O4" ~ "P2",

Genotype == "R204" ~ "Progeny"))

comp\_1A <- comp\_1A %>%

mutate(Generation2 = case\_when(

Generation == "P1" ~ "Parent",

Generation == "P2" ~ "Parent",

Generation == "Progeny" ~ "Progeny"))

comp\_1A <- comp\_1A[, c(1:2, 37:38, 3:36)]

View(comp\_1A)

#2A 29

comp\_2A\_PH29 <- DATA[DATA$Genotype %in% c("J3", "O1", "O1J3\_PH29"), ]

comp\_2A\_PH29 <- comp\_2A\_PH29 %>%

mutate(Generation = case\_when(

Genotype == "O1" ~ "P1",

Genotype == "J3" ~ "P2",

Genotype == "O1J3\_PH29" ~ "Progeny"))

comp\_2A\_PH29 <- comp\_2A\_PH29 %>%

mutate(Generation2 = case\_when(

Generation == "P1" ~ "Parent",

Generation == "P2" ~ "Parent",

Generation == "Progeny" ~ "Progeny"))

comp\_2A\_PH29 <- comp\_2A\_PH29[, c(1:2, 37:38, 3:36)]

View(comp\_2A\_PH29)

#2A 47

comp\_2A\_PH47 <- DATA[DATA$Genotype %in% c("J3", "O1", "O1J3\_PH47"), ]

comp\_2A\_PH47 <- comp\_2A\_PH47 %>%

mutate(Generation = case\_when(

Genotype == "O1" ~ "P1",

Genotype == "J3" ~ "P2",

Genotype == "O1J3\_PH47" ~ "Progeny"))

comp\_2A\_PH47 <- comp\_2A\_PH47 %>%

mutate(Generation2 = case\_when(

Generation == "P1" ~ "Parent",

Generation == "P2" ~ "Parent",

Generation == "Progeny" ~ "Progeny"))

comp\_2A\_PH47 <- comp\_2A\_PH47[, c(1:2, 37:38, 3:36)]

View(comp\_2A\_PH47)

#2B 22

comp\_2B\_PH22 <- DATA[DATA$Genotype %in% c("N5", "C2", "J1", "N5C2J1\_PH22"), ]

comp\_2B\_PH22 <- comp\_2B\_PH22 %>%

mutate(Generation = case\_when(

Genotype == "N5" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N5C2J1\_PH22" ~ "Progeny"))

comp\_2B\_PH22 <- comp\_2B\_PH22 %>%

mutate(Generation2 = case\_when(

Generation == "P1" ~ "Parent",

Generation == "P2" ~ "Parent",

Generation == "P3" ~ "Parent",

Generation == "Progeny" ~ "Progeny"))

comp\_2B\_PH22 <- comp\_2B\_PH22[, c(1:2, 37:38, 3:36)]

View(comp\_2B\_PH22)

#2B 42

comp\_2B\_PH42 <- DATA[DATA$Genotype %in% c("N5", "C2", "J1", "N5C2J1\_PH42"), ]

comp\_2B\_PH42 <- comp\_2B\_PH42 %>%

mutate(Generation = case\_when(

Genotype == "N5" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N5C2J1\_PH42" ~ "Progeny"))

comp\_2B\_PH42 <- comp\_2B\_PH42 %>%

mutate(Generation2 = case\_when(

Generation == "P1" ~ "Parent",

Generation == "P2" ~ "Parent",

Generation == "P3" ~ "Parent",

Generation == "Progeny" ~ "Progeny"))

comp\_2B\_PH42 <- comp\_2B\_PH42[, c(1:2, 37:38, 3:36)]

View(comp\_2B\_PH42)

#2C 23

comp\_2C\_PH23 <- DATA[DATA$Genotype %in% c("N5", "C2", "J2", "N5C2J2\_PH23"), ]

comp\_2C\_PH23 <- comp\_2C\_PH23 %>%

mutate(Generation = case\_when(

Genotype == "N5" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J2" ~ "P3",

Genotype == "N5C2J2\_PH23" ~ "Progeny"))

comp\_2C\_PH23 <- comp\_2C\_PH23 %>%

mutate(Generation2 = case\_when(

Generation == "P1" ~ "Parent",

Generation == "P2" ~ "Parent",

Generation == "P3" ~ "Parent",

Generation == "Progeny" ~ "Progeny"))

comp\_2C\_PH23 <- comp\_2C\_PH23[, c(1:2, 37:38, 3:36)]

View(comp\_2C\_PH23)

#2C 30

comp\_2C\_PH30 <- DATA[DATA$Genotype %in% c("N5", "C2", "J2", "N5C2J2\_PH30"), ]

comp\_2C\_PH30 <- comp\_2C\_PH30 %>%

mutate(Generation = case\_when(

Genotype == "N5" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J2" ~ "P3",

Genotype == "N5C2J2\_PH30" ~ "Progeny"))

comp\_2C\_PH30 <- comp\_2C\_PH30[, c(1:2, 37, 3:36)]

View(comp\_2C\_PH30)

#2D 24

comp\_2D\_PH24 <- DATA[DATA$Genotype %in% c("N6","C2","J2","N6C2J2\_PH24"), ]

comp\_2D\_PH24 <- comp\_2D\_PH24 %>%

mutate(Generation = case\_when(

Genotype == "N6" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J2" ~ "P3",

Genotype == "N6C2J2\_PH24" ~ "Progeny"))

comp\_2D\_PH24 <- comp\_2D\_PH24[, c(1:2, 37, 3:36)]

View(comp\_2D\_PH24)

#2D 35

comp\_2D\_PH35 <- DATA[DATA$Genotype %in% c("N6","C2","J2","N6C2J2\_PH35"), ]

comp\_2D\_PH35 <- comp\_2D\_PH35 %>%

mutate(Generation = case\_when(

Genotype == "N6" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J2" ~ "P3",

Genotype == "N6C2J2\_PH35" ~ "Progeny"))

comp\_2D\_PH35 <- comp\_2D\_PH35[, c(1:2, 37, 3:36)]

View(comp\_2D\_PH35)

#2D 44

comp\_2D\_PH44 <- DATA[DATA$Genotype %in% c("N6","C2","J2", "N6C2J2\_PH44"), ]

comp\_2D\_PH44 <- comp\_2D\_PH44 %>%

mutate(Generation = case\_when(

Genotype == "N6" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J2" ~ "P3",

Genotype == "N6C2J2\_PH44" ~ "Progeny"))

comp\_2D\_PH44 <- comp\_2D\_PH44[, c(1:2, 37, 3:36)]

View(comp\_2D\_PH44)

# 2D 48

comp\_2D\_PH48 <- DATA[DATA$Genotype %in% c("N6","C2","J2", "N6C2J2\_PH48"), ]

comp\_2D\_PH48 <- comp\_2D\_PH48 %>%

mutate(Generation = case\_when(

Genotype == "N6" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J2" ~ "P3",

Genotype == "N6C2J2\_PH48" ~ "Progeny"))

comp\_2D\_PH48 <- comp\_2D\_PH48[, c(1:2, 37, 3:36)]

View(comp\_2D\_PH48)

# 2E 25

comp\_2E\_PH25 <- DATA[DATA$Genotype %in% c("N1", "C1", "J1", "N1C1J1\_PH25"), ]

comp\_2E\_PH25 <- comp\_2E\_PH25 %>%

mutate(Generation = case\_when(

Genotype == "N1" ~ "P1",

Genotype == "C1" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N1C1J1\_PH25" ~ "Progeny"))

comp\_2E\_PH25 <- comp\_2E\_PH25[, c(1:2, 37, 3:36)]

View(comp\_2E\_PH25)

# 2E 36

comp\_2E\_PH36 <- DATA[DATA$Genotype %in% c("N1", "C1", "J1", "N1C1J1\_PH36"), ]

comp\_2E\_PH36 <- comp\_2E\_PH36 %>%

mutate(Generation = case\_when(

Genotype == "N1" ~ "P1",

Genotype == "C1" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N1C1J1\_PH36" ~ "Progeny"))

comp\_2E\_PH36 <- comp\_2E\_PH36[, c(1:2, 37, 3:36)]

View(comp\_2E\_PH36)

# 2E 31

comp\_2E\_PH31 <- DATA[DATA$Genotype %in% c("N1", "C1", "J1", "N1C1J1\_PH31"), ]

comp\_2E\_PH31 <- comp\_2E\_PH31 %>%

mutate(Generation = case\_when(

Genotype == "N1" ~ "P1",

Genotype == "C1" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N1C1J1\_PH31" ~ "Progeny"))

comp\_2E\_PH31 <- comp\_2E\_PH31[, c(1:2, 37, 3:36)]

View(comp\_2E\_PH31)

# 2E 39

comp\_2E\_PH39 <- DATA[DATA$Genotype %in% c("N1", "C1", "J1", "N1C1J1\_PH39"), ]

comp\_2E\_PH39 <- comp\_2E\_PH39 %>%

mutate(Generation = case\_when(

Genotype == "N1" ~ "P1",

Genotype == "C1" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N1C1J1\_PH39" ~ "Progeny"))

comp\_2E\_PH39 <- comp\_2E\_PH39[, c(1:2, 37, 3:36)]

View(comp\_2E\_PH39)

# 2F 26

comp\_2F\_PH26 <- DATA[DATA$Genotype %in% c("N1", "C2", "J1", "N1C2J1\_PH26"), ]

comp\_2F\_PH26 <- comp\_2F\_PH26 %>%

mutate(Generation = case\_when(

Genotype == "N1" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N1C2J1\_PH26" ~ "Progeny"))

comp\_2F\_PH26 <- comp\_2F\_PH26[, c(1:2, 37, 3:36)]

View(comp\_2F\_PH26)

# 2F 40

comp\_2F\_PH40 <- DATA[DATA$Genotype %in% c("N1", "C2", "J1", "N1C2J1\_PH40"), ]

comp\_2F\_PH40 <- comp\_2F\_PH40 %>%

mutate(Generation = case\_when(

Genotype == "N1" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N1C2J1\_PH40" ~ "Progeny"))

comp\_2F\_PH40 <- comp\_2F\_PH40[, c(1:2, 37, 3:36)]

View(comp\_2F\_PH40)

#2G 27

comp\_2G\_PH27 <- DATA[DATA$Genotype %in% c("N4","C2","J1","N4C2J1\_PH27"), ]

comp\_2G\_PH27 <- comp\_2G\_PH27 %>%

mutate(Generation = case\_when(

Genotype == "N4" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N4C2J1\_PH27" ~ "Progeny"))

comp\_2G\_PH27 <- comp\_2G\_PH27[, c(1:2, 37, 3:36)]

View(comp\_2G\_PH27)

#2G 45

comp\_2G\_PH45 <- DATA[DATA$Genotype %in% c("N4","C2","J1","N4C2J1\_PH45"), ]

comp\_2G\_PH45 <- comp\_2G\_PH45 %>%

mutate(Generation = case\_when(

Genotype == "N4" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N4C2J1\_PH45" ~ "Progeny"))

comp\_2G\_PH45 <- comp\_2G\_PH45[, c(1:2, 37, 3:36)]

View(comp\_2G\_PH45)

#2H

comp\_2H <- DATA[DATA$Genotype %in% c("N7","C1","J1","N7C1J1"), ]

comp\_2H <- comp\_2H %>%

mutate(Generation = case\_when(

Genotype == "N7" ~ "P1",

Genotype == "C1" ~ "P2",

Genotype == "J1" ~ "P3",

Genotype == "N7C1J1" ~ "Progeny"))

comp\_2H <- comp\_2H[, c(1:2, 37, 3:36)]

View(comp\_2H)

#3A

comp\_3A <- DATA[DATA$Genotype %in% c("J1", "C1", "J1C1"), ]

comp\_3A <- comp\_3A %>%

mutate(Generation = case\_when(

Genotype == "J1" ~ "P1",

Genotype == "C1" ~ "P2",

Genotype == "J1C1" ~ "Progeny"))

comp\_3A <- comp\_3A[, c(1:2, 37, 3:36)]

View(comp\_3A)

#3B

comp\_3B <- DATA[DATA$Genotype %in% c("J1", "C2", "J1C2"), ]

comp\_3B <- comp\_3B %>%

mutate(Generation = case\_when(

Genotype == "J1" ~ "P1",

Genotype == "C2" ~ "P2",

Genotype == "J1C2" ~ "Progeny"))

comp\_3B <- comp\_3B[, c(1:2, 37, 3:36)]

View(comp\_3B)

#4A

comp\_4A <- DATA[DATA$Genotype %in% c("N5C2J2\_PH30", "N1C1J1\_PH31", "N1C1J1.N5C2J2"), ]

comp\_4A <- comp\_4A %>%

mutate(Generation = case\_when(

Genotype == "N1C1J1\_PH31" ~ "P1",

Genotype == "N5C2J2\_PH30" ~ "P2",

Genotype == "N1C1J1.N5C2J2" ~ "Progeny"))

comp\_4A <- comp\_4A[, c(1:2, 37, 3:36)]

View(comp\_4A)

#4B

comp\_4B <- DATA[DATA$Genotype %in% c("N7C1J1", "N6C2J2\_PH24", "N6C2J2.N7C1J1"), ]

comp\_4B <- comp\_4B %>%

mutate(Generation = case\_when(

Genotype == "N6C2J2\_PH24" ~ "P1",

Genotype == "N7C1J1" ~ "P2",

Genotype == "N6C2J2.N7C1J1" ~ "Progeny"))

comp\_4B <- comp\_4B[, c(1:2, 37, 3:36)]

View(comp\_4B)

#4C

comp\_4C <- DATA[DATA$Genotype %in% c("N7C1J1", "N5C2J2\_PH23", "N5C2J2.N7C1J1"), ]

comp\_4C <- comp\_4C %>%

mutate(Generation = case\_when(

Genotype == "N5C2J2\_PH23" ~ "P1",

Genotype == "N7C1J1" ~ "P2",

Genotype == "N5C2J2.N7C1J1" ~ "Progeny"))

comp\_4C <- comp\_4C[, c(1:2, 37, 3:36)]

View(comp\_4C)

#4D

comp\_4D <- DATA[DATA$Genotype %in% c("N6C2J2\_PH35", "N1C1J1\_PH36", "N1C1J1. N6C2J2"), ]

comp\_4D <- comp\_4D %>%

mutate(Generation = case\_when(

Genotype == "N1C1J1\_PH36" ~ "P1",

Genotype == "N6C2J2\_PH35" ~ "P2",

Genotype == "N1C1J1. N6C2J2" ~ "Progeny"))

comp\_4D <- comp\_4D[, c(1:2, 37, 3:36)]

View(comp\_4D)

#4E

comp\_4E <- DATA[DATA$Genotype %in% c("N5C2J2\_PH30", "N6C2J2\_PH24", "N5C2J2.N6C2J2"), ]

comp\_4E <- comp\_4E %>%

mutate(Generation = case\_when(

Genotype == "N5C2J2\_PH30" ~ "P1",

Genotype == "N6C2J2\_PH24" ~ "P2",

Genotype == "N5C2J2.N6C2J2" ~ "Progeny"))

comp\_4E <- comp\_4E[, c(1:2, 37, 3:36)]

View(comp\_4E)

#4F

comp\_4F <- DATA[DATA$Genotype %in% c("N1C1J1\_PH39", "N1C2J1\_PH40", "N1C1J1.N1C2J1"), ]

comp\_4F <- comp\_4F %>%

mutate(Generation = case\_when(

Genotype == "N1C1J1\_PH39" ~ "P1",

Genotype == "N1C2J1\_PH40" ~ "P2",

Genotype == "N1C1J1.N1C2J1" ~ "Progeny"))

comp\_4F <- comp\_4F[, c(1:2, 37, 3:36)]

View(comp\_4F)

#4G ###KORRIGIEREN ###DANI

comp\_4G <- DATA[DATA$Genotype %in% c("N5C2J1\_PH42", "N5C2J2\_PH30", "N5C2J2.N5C2J2"), ]

comp\_4G <- comp\_4G %>%

mutate(Generation = case\_when(

Genotype == "N5C2J1\_PH42" ~ "P1",

Genotype == "N5C2J2\_PH30" ~ "P2",

Genotype == "N5C2J2.N5C2J2" ~ "Progeny"))

comp\_4G <- comp\_4G[, c(1:2, 37, 3:36)]

View(comp\_4G)

#4H

comp\_4H <- DATA[DATA$Genotype %in% c("N6C2J2\_PH44", "N4C2J1\_PH45", "N6C2J2.N4C2J1"), ]

comp\_4H <- comp\_4H %>%

mutate(Generation = case\_when(

Genotype == "N6C2J2\_PH44" ~ "P1",

Genotype == "N4C2J1\_PH45" ~ "P2",

Genotype == "N6C2J2.N4C2J1" ~ "Progeny"))

comp\_4H <- comp\_4H[, c(1:2, 37, 3:36)]

View(comp\_4H)

#4I

comp\_4I <- DATA[DATA$Genotype %in% c("O1J3\_PH47", "N6C2J2\_PH48", "N6C2J2.O1J3"), ]

comp\_4I <- comp\_4I %>%

mutate(Generation = case\_when(

Genotype == "N6C2J2\_PH48" ~ "P1",

Genotype == "O1J3\_PH47" ~ "P2",

Genotype == "N6C2J2.O1J3" ~ "Progeny"))

comp\_4I <- comp\_4I[, c(1:2, 37, 3:36)]

View(comp\_4I)

**Workshop (18.-30.4.)**

Ich tue gerade genau das richtige! Ich erwerbe jeden Tag neues Wissen und neue Fähigkeiten. Gleichzeitig lege ich den Fokus auf konkrete Skills, die der Auswertung meiner Daten dienen. So beherrsche ich meine Herausforderungen (zum Beispiel R) mit der Zeit immer besser. Das ist eine gut durchdachte, zielorientierte und langfristig gesehen effiziente Vorgehensweise.

In ein paar Jahren, wenn ich die neuen Themen und Methoden beherrsche, werde ich auf diese Zeit als eine Phase des Empowerments zurückblicken, in der ich mir wichtige Kenntnisse und Fähigkeiten erarbeitet habe. Diese Phase muss Annaliese auch gehabt haben, in der sie sich alles draufgeschafft hat.

- Make analysis plan (see below)

- What do I want to produce? How is Heterosis commonly given?

- mornings: read two books

- afternoons: transfer what I have learned to my analyses

- end of first week: send current state to Kevin, Fei and R community to get help

- end of second week: have brand new exciting results revealing new insights

Check: do I need sex of parents? Merge BBCH data?

**Analysis plan**

Ein gut strukturiertes separates Dokument mit den Hypothesen, ergänzenden Fragestellungen und geplanten statistischen Analysen führen. Informationen aus Annalieses Excel Sheet nutzen. Was will ich wissen? Warum welcher statistische Test? Was ist jeweils die übliche bzw. beste Darstellung (tabellarisch und graphisch)? Zweck: Langfristige Orientierung.

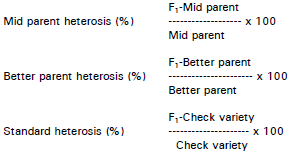
Beim Programmieren Schritte planen, in R übertragen und Schritt für Schritt lösen.

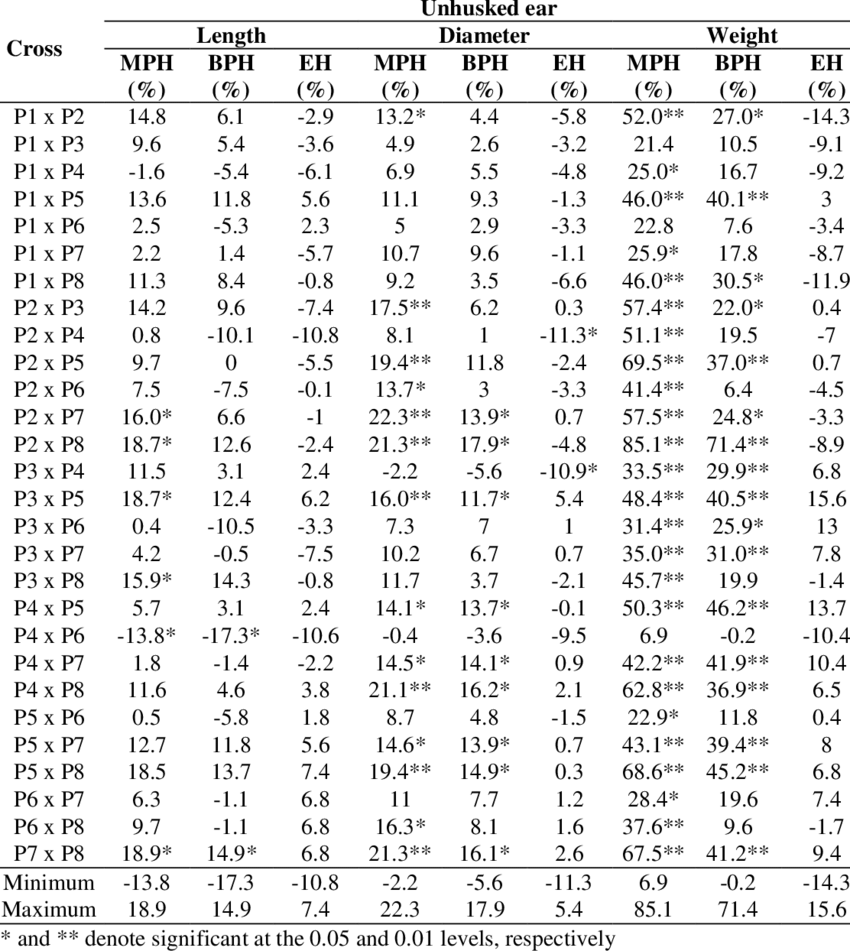
Ggf. auch Quellen zu bestimmten Analysen angeben.

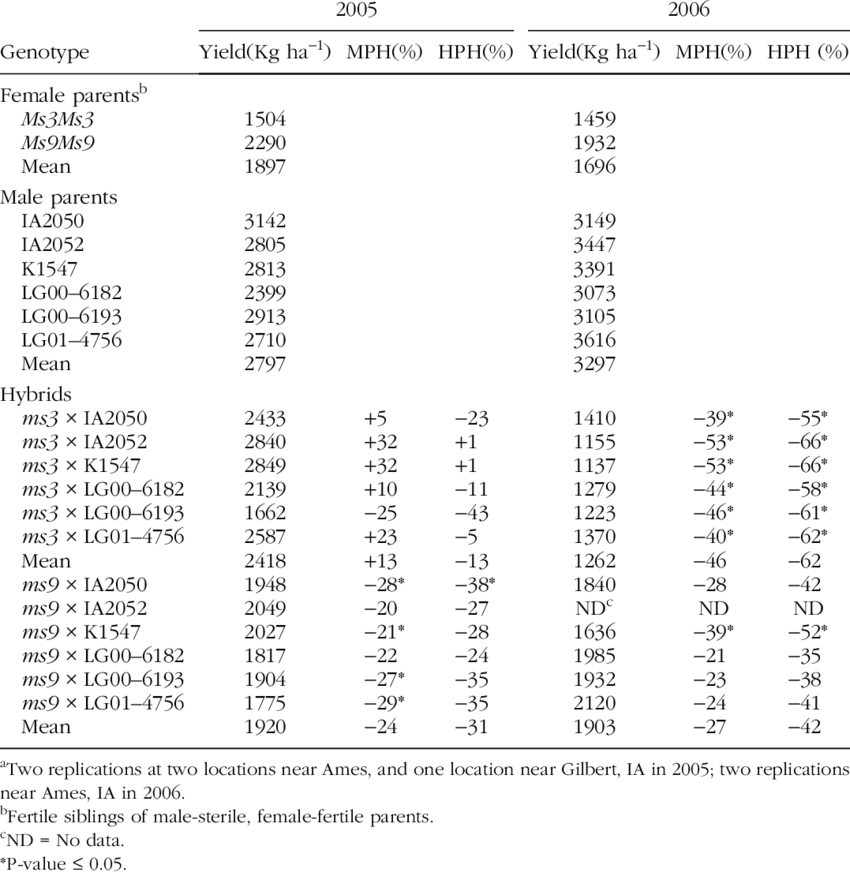
How is MPH and BPH commonly given?

+/- absoluter oder relativer Unterschied mit Signifikanzniveau \*\*\*

**Wie wird Heterosis üblicherweise angegeben?**







**27-04-22**

1. Was will ich haben?

- Mittelwerte, absolute und relative Differenz

- ANOVA und T-Test

- Assumptions

- MPH uns BPH nach Formel (= Prozentuale Veränderung von MPV)

2. Wie produziere ich dieses Ergebnis? Was funktioniert schon? Auf **map** functions festlegen, apply und loop weglassen. Ich muss ja nur einen funktionierenden Code schreiben.

3. Fragen:

- Ist es besser eine vorhandene Tabelle (nested, siehe Video Wickham) um Spalte mit extrahierten p-Werten zu ergänzen oder wie gehabt p-Werte als separaten Output zu erzeugen und dranzuflanschen?

- Wie erzeuge ich die Spalte Generation2 für parents mean?

4. Zu erst diese Kapitel lesen:

- Iteration

- many Models

- subsetting

- Conditions

- functions

- Functionals

1. Listen von den Vergleichsgruppen erstellen

2. Nesting nach Listenelementen (parent progeny Sets)

3. extra Spalte Generation erstellen

4. daraus Ergebnistabelle gleichen Formats erstellen

**29.04.22 Ergebnisse First level analyses erzeugen**

#######################################################################

##New approach, partly from epirhandbook.com #29.4.22##########################

# Results are saved as a list

levene.tests <- c1A %>%

select\_if(is.numeric) %>%

map(~ leveneTest(.x ~ c1A$Generation2))

levene.tests

names(levene.tests)

Levene\_TTest\_Out <- map(names(levene.tests), function(trait) {

i1 <- levene.tests[[trait]]$"Pr(>F)"[1] >= 0.05

t.test(c1A[[trait]]~ c1A$Generation2, var.equal = i1)

})

Levene\_TTest\_Out

# to extract not only p-values, but also means for each group, we would need to

# extract the element estimate, but this actually contains two elements within it

# (mean in group f and mean in group m). So, it cannot be simplified into a vector

# with map\_chr() or map\_dbl(). Instead, we use map() within tibble() to

# create a column of class list within the tibble! Yes, this is possible!

#

# Below, we pass the tibble to unnest\_wider() specifying the tibble’s means

# column (which is a nested list). The result is that means is replaced by two

# new columns, each reflecting the two elements that were previously in each

# means cell.

T.Test\_table\_c1A <- Levene\_TTest\_Out %>%

{tibble(

traits = names(levene.tests),

p = map\_dbl(., "p.value"),

means = map(., "estimate")

)} %>%

unnest\_wider(means)

T.Test\_table\_c1A

# change colnames##

colnames(T.Test\_table\_c1A) <- c("traits", "p", "mid\_parent\_value", "progeny\_mean")

T.Test\_table\_c1A

#neue spalten einfügen

T.Test\_Results\_c1A <- T.Test\_table\_c1A %>%

mutate(Diff\_MPV = progeny\_mean - mid\_parent\_value,

MPH = ((progeny\_mean - mid\_parent\_value) / mid\_parent\_value) \* 100)

T.Test\_Results\_c1A

**Answer Annaliese results, 11.5.22**

Hi Helen,

so I have had a look, but it is for sure not so easy to work out what is happening with all the data :P Still, having too much data is a bit of a luxury problem! The summary tables at the ends are good, you mean "Better Parent Heterosis" and "Mean Parent Heterosis" with BPH and MPH I assume, so the progeny compared to either the better parent (or worse parent I assume) for BPH, then to the mid-parent mean for MPH? Yes, I calculated Mid Parent Heterosis and Better Parent Heterosis using the common formula giving % deviation from the respective parent or mid parent value.

So this is for sure what we are interested in for each trait, but the BPH tables are hard to read because you included all the values for each parent compared to the progeny, not just the "best" or "worst" parent. True!

Even though it will add extra columns I would recommend having the actual trait values for each of P1, P2 and the hybrid in the final summary tables and not the percentages, because it took me a while to understand you had already converted it into percentages, and for the traits it also helps if we have the real number (with known units also if required, so e.g. days to flowering) also for comparison. Percentages are not so interesting at this stage as which traits are significantly different between progeny and parents and in which direction. Alternatively just include the difference in real numbers instead of percentages instead of adding extra columns. No, problem! I have the values and I can easily include them. I just wanted to reduce the size of the table. That is why I asked you last week which single value you would prefer, if I show only one per comparison set and trait.

My recommendation as the next step is to start summarising in text form what the results are telling us in terms of significant comparisons. This might also involve grouping some sets of measurements into the same "trait", for example growth rate or flowering time measurements, just for summary purposes. This might look something like as follows:

"Across the seven NCJ allohexaploid genotype combinations (17 progeny sets), most progeny showed significantly lower seed fertility compared to the mid-parent mean (16/17 progeny sets showed significantly reduced seed weight per 10 pods, and 12/17 progeny sets showed significantly lower seed number per pod). Only one progeny set showed similar seed fertiliity to the mid-parent mean for both measured seed traits. Days to flowering in the NCJ allohexaploids was significantly delayed across all genotype combinations (although not in all progeny sets), on average by XX days across all genotypes."  Please note I am not actually sure if these are legitimate (I think days to flowering might actually be faster??), I'm trying to go off the tables but it's a bit difficult for me to try and turn percentages back into real values then assess this... maybe just the numbers (e.g. 5 for an additional 5 days to flower relative to best parent or mid-parent value) instead of percentages here would be fine. In principle better parent heterosis is more interesting (faster or slower, taller or smaller than both parents) but this is also a bit tricky for me to read off the tables, I have to put a lot of mental effort into understanding these at the moment 😝 Oh, sorry!;) So, in the new table I will give the means of parents and progeny (or MPV) and the absolute differences. And I will summarise the results per comparison group in a text. Thanks for the example!

Another thing which would be super helpful would be to start on the methods and have already a written section which describes what each trait means - I don't know what TGW is for instance, I thought this was thousand seed weight then realised that should be TSW... We should probably also work out which of the digital phenotypes can be grouped together or which were most useful etc.. Which traits are correlated with each other across the dataset? I assume you are right and TGW means “thousand grain weight”. So, this methods section should be written by Kevin and Fiona, right? There is also some information on imaging traits in the documents Kevin provided. I’ll have a look.

Yes, the correlation analysis for the traits within each parent-progeny set is going to be the next step.

Best wishes,

Annaliese

See additional comments in blue:

Am 06/05/2022 um 14:24 schrieb [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de):

Hi Annaliese,

hi Dani, Henrik and Anne, if you are also interested,

if you want to take a look, I have saved the first level analyses on Share/EPPN Data Analyses. **For each of the four comparison groups there are two files holding the results for MPH and BPH (%) assessed by a Welch’s T-Test. The last sheet on the right (called MPH or BPH) contains the finals results.** Bold values are significant at least at the 5% level. (Subgroup 3A is currently missing. The t-test didn’t work for this set, probably because there were only three individuals for one of the parents.)

There are also results for **comparison 1X**, shown as boxplots with significances based on a t-test. Up to now, this is only for the complete group 1A, not for B and C.

This is not the complete 1X group - it makes no sense to just have R2O4 plus R2 and O4 and the *B. napus* but not the other two synthetics or their parents, I was really confused about this! If we want to compare all synthetics to all parents plus the *B. napus* everything has to be in the same plot. I think the 1B and 1C comparisons from the original file are missing still, right? They are incomplete. I think one parent is missing in both sets.

I hope this way of presenting the results is ok for you. I think most traits are self-explanatory. The traits assessed by imaging analyses (and their codes) are given in the screenshot below.

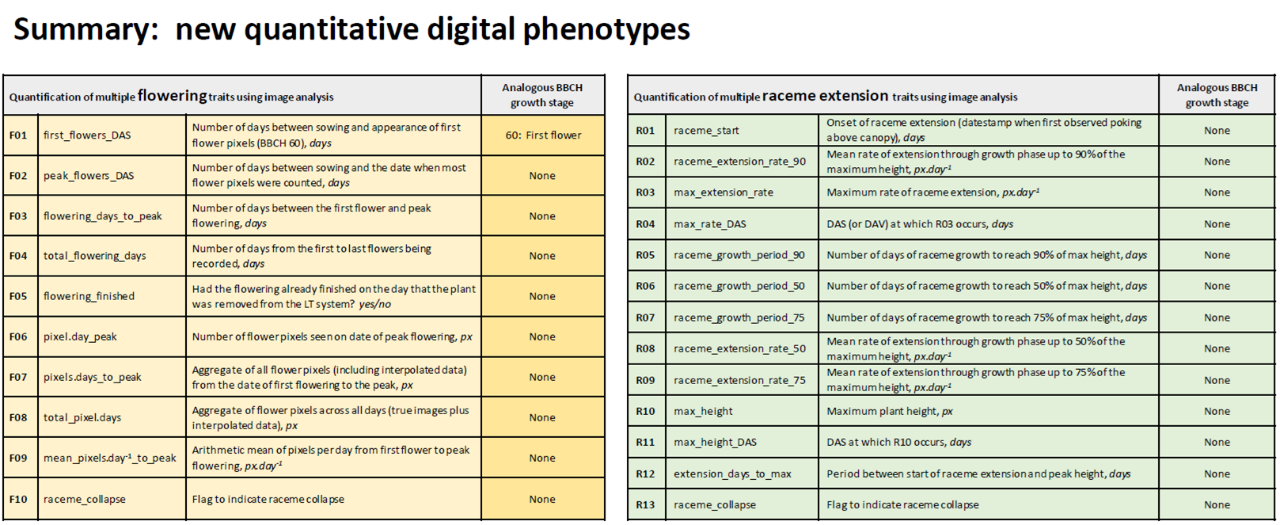
The boxplots with points look great - I like this presentation method for comparisons in the data! I don't really like the way of indicating significance by drawing lines above, this is a bit hard to read - I think the more common practice of just adding the letters above the plots to indicate which lines are siginificantly different from each other is better, although this is only a minor point.

If you have any questions or comments, please let me know. Can we discuss these results next week or the following? Thanks!

Have a nice weekend!

Best wishes

Helen



**summary**

* Edit summary tables:
  + the actual trait values for each of **P1, P2 and the hybrid** in the final summary tables
  + **Alternatively just include the difference** in real numbers instead of percentages instead of adding extra columns. separately.
  + **include 3A!**
  + In principle better parent heterosis is more interesting
  + man könnte auch noch MPV erstellen
* write text
  + start summarising in text form what the results are telling us in terms of significant comparisons.
  + This might also involve grouping some sets of measurements into the same "trait", for example growth rate or flowering time measurements, just for summary purposes. see example.
* start on the methods section
  + which describes what each trait means.
  + check the early Excel tables we received first to see if we have more detailed information on the traits, and if not contact Fiona et al. for additional explanations.
* work out which of the digital phenotypes can be grouped together or which were most useful etc.. (for methods section or evaluation? Egal, Kevin fragen!)
* Comparison 1X:
  + we want to compare all synthetics to all parents plus the B. napus.
  + The boxplots with points look great - I like this presentation method for comparisons in the data!
  + the more common practice of just adding the letters above the plots to indicate which lines are significantly different from each other
* Correlation analyses:
  + Which traits are correlated with each other across the dataset?

**13.5.22**

1. Was bedeuten die Traits? Wie sind Sie erhoben worden? Welche kann ich im Text zusammenfassen? In Kevins Unterlagen schauen und ihm Fionas Mails und Exel Dateien. Fehlende Informationen erfragen.

2. absolute Unterschiede zu parents und MPH berechnen (Gruppen eins und zwei)

3. Ergebnisse für jede der vier Gruppen in Textform zusammenfassen. Bis dahin liegen sicher auch die Antworten von Kevin und Fiona vor.

**16.5.22**

1. Calculate absolute differences
2. Summarise results per group
   1. Fokus: Text mit Fokus auf die Eigenschaften der Nachkommen gegenüber Ihren Eltern formulieren
   2. start each section (growth / fertility) with summarising phrase
   3. Wichtigste Parameter auswählen und in logischen Zusammenhang setzen (dann für alle 4 Gruppen so schreiben)
   4. Liegt die Mehrheit der Nachkommenlinien über oder unter dem Durchschnitt der Eltern?
   5. Wieviele Nachkommen liegen sogar unter/über dem höchsten/niedrigsten Elternteil?
   6. Unterschiede in absoluten Zahlen angeben
   7. wichtig: Absolute Werte statt % angeben. Nur mit SIGNIFIKANTEN Unterschieden arbeiten.
   8. Confirm/reject hypotheses.
   9. see Annalieses example: flowering delayed, grow faster, produce more biomass, show increased fertility, don’t stick to the parameters, but close to data
   10. avoid too many substantives
   11. Bei hoher Variabilität: “have high segregation for a certain trait”
3. X-analyses: redo with complete sets and letters for significances. Summarise in text format.

Hi Helen,

I also find you very easy to work with and very productive on very little feedback! :) Responses below in blue:

Am 09/06/2022 um 18:25 schrieb [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de):

Hi Annaliese,

Thanks for your helpful feedback! Working with you is very efficient😊. I will go through it in detail tomorrow.

Missing data analyses:

1. I’m currently working on the X and Y comparisons to figure out whether the hybrids are more similar to each other or to their parents. See attached figure. The genotypes need to be sorted in a different order, but if you consider this useful I can create these figures for all X and Y comparisons.

I think this general type of figure is good. We might need to split it into boxes/parts that each show individual hybrid progeny sets in comparison to their exact parent genotypes though, because I think it will be tricky to see the patterns otherwise...

1. Furthermore we were planning to do correlation analyses to check the correlation between traits.

Yes, good!

1. In the beginning you suggested to do a PCA. Do you still think this would be useful?

Yes, this is similar to the correlation analyses, but can generate a nice phenotype vectors figure which may be helpful to visualise the relationships between the traits.

1. Second, third, fourth level analyses with corresponding figures?

Ah, right... So this means here we would e.g. take the parent *B. napus*, *B. juncea* and *B. carinata* genotypes, their derived hexaploids, then also the F1 hybrids between the hexaploids, and try and show a trend? This would be interesting I think to see how some traits are changing based on this :) I don't think we necessarily have to combine everything together, but the NCJ x NCJ hexaploids where we have the full set of parent genotypes could be interesting and clear, we'd have to see what this looks like as a figure.

Yes, all reported differences are significant. Otherwise I wouldn’t call them “differences”. But I can state this more clearly.

You said ANOVA is robust to violations of the normality assumption. Are T-Test robust, too? If not, I have to check this for all reported results. (This is in an separate document and I didn’t manage to create a conditional analysis code.) Unequal variances are no problem, since I used the Welch’s T-Test.

There is some literature on this which you may want to check since I'm not an expert, but as far as I know t-tests and ANOVAs use identical underlying statistics, ANOVA is just able to include more than one comparison, so I imagine it's similar.

In general, we have to decide how to deal with differences between data from hand measurements and image analyses. Maybe we can also discuss this with Kevin and Fiona.

Yes, I was wondering if this is also a result - it's not one I find particularly thrilling, since it's more technical, but maybe we need to include it if it's not already well-established in the literature. If the pipelines are already in use/published, then I think we can just reference this in the methods, and maybe discuss in the Discussion section if we want.

Have you heard anything from Fiona recently? She hasn’t answered my questions about the methods section, e.g. on units.

Haven't heard anything, sorry!

In the next journal club I would like to give a short presentation on R courses and resources that I recommend and discuss a book chapter on debugging. Is that ok?

Yes, sounds good!

Best wishes

Helen

**Von:** Annaliese Mason [<annaliese.mason@uni-bonn.de>](mailto:annaliese.mason@uni-bonn.de)   
**Gesendet:** Donnerstag, 9. Juni 2022 15:16  
**An:** [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de)  
**Betreff:** Re: EPPN: parent-progeny comparisons in text form

Hi Helen,

looks fantastic, great work! I went through in detail and made some suggestions (attached).

Are there any additional results which we are missing that need to be processed etc., or is this now everything already for the results section? If this is the case then I think we only need maybe one last paragraph comparing across the different types of progeny (see attached), then this is already finished except for deciding which results are the most interesting and then choosing what figures we want in the main text accordingly :p

Best wishes,

Annaliese

Am 08/06/2022 um 10:34 schrieb [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de):

Hi Annaliese,

This is the current version of the text (with a few edits).

Best wishes

Helen

**Von:** Annaliese Mason [<annaliese.mason@uni-bonn.de>](mailto:annaliese.mason@uni-bonn.de)   
**Gesendet:** Dienstag, 7. Juni 2022 14:01  
**An:** [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de)  
**Betreff:** Re: EPPN: parent-progeny comparisons in text form

Hi Helen,

sorry, I was sick last week so didn't manage to meet my self-imposed deadline of before you got back for looking at this! I will do my best in the next few days :) I did open it already and have a look; one suggestion in case you want to move forward in the meantime would be to identify for each major result which figure would be the best to show this data, and to start building then the results section figures in this way :)

Best wishes,

Annaliese

Am 07/06/2022 um 11:34 schrieb [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de):

Hi Annaliese,

Thanks for your feedback! Did you have the chance to look at the results in detail, yet? We could also meet and talk about them in person, if you have time.

Best wishes

Helen

**Von:** Annaliese Mason [<annaliese.mason@uni-bonn.de>](mailto:annaliese.mason@uni-bonn.de)   
**Gesendet:** Freitag, 20. Mai 2022 15:31  
**An:** [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de)  
**Betreff:** Re: EPPN: parent-progeny comparisons in text form

Hi Helen,

that looks fantastic, excellent work! I will have a look in detail before you get back from holiay then, and will have a look also at the updated Excel tables to see if I have any additional questions or suggestions :)

The holiday was super nice but rather exhausting, I'm glad it's the weekend :p

All the best!

Annaliese

Am 20/05/2022 um 14:39 schrieb [hbehn@uni-bonn.de](mailto:hbehn@uni-bonn.de):

Hi Annaliese,

So, I’ve summarised the parent-progeny comparisons in text form (attached). Some interesting differences, but you know the genotypes better than me.

I hope you had a nice short holiday on Monday and Tuesday? I’m going on holiday tomorrow. Two weeks travelling through France and the Spanish Pre-Pyrenees! 😊

See you in two weeks! All the best!

Helen

**17.6.22**

So hatte der Code funktioniert, aber die Boxplots waren in der falschen Reihenfolge:

> ## create a list of your data.frames

> comp2.4 <- list(comp\_2B.4, comp\_2C.4, comp\_2D.4, comp\_2E.4, comp\_2F.4, comp\_2G.4,

+ comp\_2H.4)

>

> # naming 7 Parent combinations in list for addressing them later on ##

> names(comp2.4) <- c("N5C2J1", "N5C2J2", "N6C2J2", "N1C1J1", "N1C2J1", "N4C2J1", "N7C1J1")

> Bound\_comp2.4 <- rbindlist(comp2.4, id = "parent\_combination")

> # create the plot

> p <- ggplot(data = Bound\_comp2.4, aes(x = Genotype,

+ y = plant\_height,

+ color = Generation)) +

+ geom\_boxplot() +

+ theme(axis.title = element\_text(size = 14),

+ legend.text = element\_text(size = 7),

+ legend.title = element\_text(size = 10),

+ axis.title.x = element\_text(size = 10),

+ axis.title.y = element\_text(size = 10),

+ axis.text.x = element\_text(size = 6),

+ axis.text.y = element\_text(size = 6)) +

+ rotate\_x\_text(45)

>

> p + facet\_wrap(~ parent\_combination, scales = "free\_x")

**7.6.22**

**To do**

* genotyp bezeichnung dani korrigieren
* check T-Test assumptions
* do all X- and Y-analyses
* ASM
  + correlation analyses
  + PCA
  + identify for each major result which figure would be the best to show this data, and to start building then the results section figures in this way :)
  + second/third level analyses: Ah, right... So this means here we would e.g. take the parent B. napus, B. juncea and B. carinata genotypes, their derived hexaploids, then also the F1 hybrids between the hexaploids, and try and show a trend? This would be interesting I think to see how some traits are changing based on this :) I don't think we necessarily have to combine everything together, but the NCJ x NCJ hexaploids where we have the full set of parent genotypes could be interesting and clear, we'd have to see what this looks like as a figure.
  + please make sure to specify each time for each result if it is significant or just a trend. I think you only reported significant results, but it needs to be 100% clear in the text.
  + I think it would be good to also report the negative results in some cases just to be extra clear: so we have three possible outcomes for each comparison, which is greater/same/less than parents. It might be clearer to report like follows: “Of the XX progeny sets, XX were significantly taller than the mid-parent mean, while the rest were not significantly different to the mid-parent mean” or specify how many were below or above.
  + **Which traits were most affected by allopolyploidisation**, and in which direction? (e.g. : was flowering time generally delayed, fertility generally decreased, and plant height increased?)
  + **What does the comparison of the novel allohexaploids to their F1 hybrids tell us about relative contributions of allelic heterosis and increase in ploidy level?** (do we see any difference in **which traits** are affected when we go from allotetraploid to allohexaploid compared to when we make allohexaploid hybrids?)
  + We might need more details at the start here because these are different types: some are between species combinations, others just NCJ x NCJ genotypes etc., and there is one is just between two lineages derived from the same parent genotypes

**QUESTIONS TO KEVIN**

* what causes these error messages:
  + Removed 11 rows containing non-finite values (stat\_signif).
  + Removed 11 rows containing missing values (geom\_point).