

Identifying the Best Giant Pumpkin Seed: A Statistical Approach

Joe Ailts, Fall 2018



2109 Zywiec, from the 1781 Zywiec

Gordon Thompson, in Langevin's early edition of "How to Grow World Class Giant Pumpkins", famously said "All you need is good seed, good soil, good weather, and good luck". Excluding soil, all are factors that are hard for the gardener to control. In the absence of the ability to control these three, the best we can do is accept the hand we're dealt and minimize risk along with maximizing upside. With luck, it is highly correlated to time invested. Maximizing the uncontrollable upside of luck means toiling away in the patch for more hours, thereby increasing your chances of a respectable fruit come

season's end. With weather, it is more about reducing risk. Building wind fences, covering pumpkins with thick blankets during hailstorms (while taking a pummeling to your own body parts in the process), and driving home over the noon hour to mist during heat waves all serve to minimize risk associated with uncontrollable weather. What about "good seed" though? Most competitive giant pumpkin growers have set themselves up for success by planting seeds from the biggest pumpkins in the world. I would classify this as mitigating risk vs maximizing upside, however...minimizing the chances you'll top out at 300lbs using off-the-shelf Atlantic Giant ho-hum seed. What are some evidence-based strategies for maximizing upside in seed selection? The vast majority of growers will simply try to emulate the most successful "Heavy Hitters" in the hobby. It is not a faulty strategy. Emulating the best in any hobby or endeavor is often a default path to one's own success. While sound logic, I propose a question- do the top 1%'ers in this hobby have a secret or unique seed selection process that enables them to maximize upside in seed genetics? I doubt they do. Rather, they'll lean on maximizing upside and minimizing risk in the other factors: weather, soil, luck/time...to ensure contest-winning outcomes in their patch year over year.

Seed selection is one of the most exciting aspects of this hobby. Growers have the entire off season to pore over myriad genetics, identify the crosses they believe will succeed in their patches, and engage in lively discussions on the message boards to defend their favorites. While a highlight of the competitive enthusiast's process, seed selection has been a highly subjective exercise rooted in gut-feelings and minimal analytics-based decision making. Assuming your goal is to grow the largest pumpkin possible, by what process would you utilize to ensure you had the absolute best genetics in the hobby?

I have long been a proponent of the "3-way Cross Weight Average" method of seed selection, which you can read a full-length article on in the member's-only section of the St. Croix Growers website. The gist of this method suggests that the biggest pumpkins come from the biggest parental lineage. It is a forward-looking, back of the envelope way to theoretically maximize upside. While I still believe this is a useful tool for screening top-potential genetics, I have also begun crunching historical data and using some rudimentary statistical analysis to determine if we can quantify maximum upside in seed selection.

Ask any competitive grower in this hobby what their choice for the "best seed" is over the 2017-2018 growing seasons and I bet 9 out of 10 will say "2145 McMullen". It is not necessarily a wrong guess, but

how do we know it is right? And that's the point here...it is a guess. The process is fraught with a multitude of biases, few of which are rooted in reliable data. No question this seed has produced an amazing collection of top-end pumpkins, including the current world record 2624 Willemijns, but is it truly, objectively, the *best* seed in the hobby? As a born skeptic, I enjoy challenging group-think conclusions and utilizing data to help unearth more quantifiable results. So that's what I did here, and the results are fascinating! Using 3 three years (2016, '17, '18) of historical data, pulled from www.bigpumpkins.com GPC pumpkin weigh off data pages, I was able to explore seed genetics and performance in ways that have not been done before.

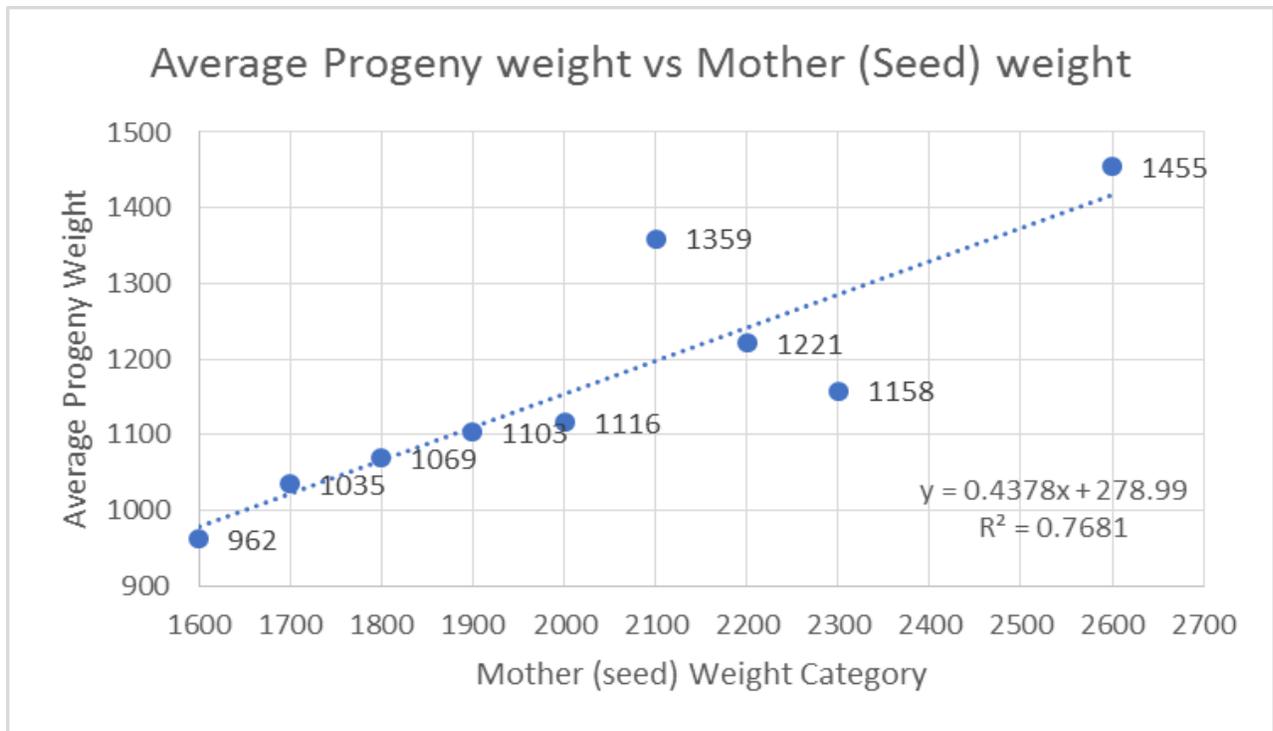
Two major outcomes to share right off the top: 2145 McMullen, while a top-performing seed, may *not* be the top performing seed in the hobby. Secondly, data is suggesting the 3-way cross weight average method has merit, as pumpkins from heaviest mothers tend to produce heavier pumpkins (who woulda thunk?!)

Let's dig in to the methods, data, and analysis. To reiterate, I am working with every GPC pumpkin entered into the BigPumpkins.com database from 2016-2018....a total of 5324 pumpkins. After clean up and removal of entries that did not have a female pumpkin listed, I then sorted all entries by their maternal parentage. This permitted me to then calculate average progeny weight outcomes for each seed planted. To make the final list, I set an arbitrary threshold of 8 entries per seed. The number 8 does not have any sort of statistical significance, I choose it in the interest of populating a more complete list. Further, statistical relevance becomes more trustworthy once you reach 30+ entries. Only 5 seeds in the last 3 years have resulted in 30 or more official GPC entries. As such, a purely statistically significant evaluation of top seed genetics using just these 5 seed lines would be extremely limited in its scope. Lastly, here is the obligatory disclaimer regarding interpretation of this data. By no means is this a definitive statement on the best seeds in the hobby. This is a rudimentary quantification of average fruit weight from various seed lines. 1001 factors can and have influenced the numbers in the charts below. These include and are not limited to: number of years the seed has been available, number of seeds harvested, regional distribution of those seeds, grower bias within the patch towards a preferred seed, and pumpkins weighed at non-GPC events that are not included in the data set to name a few. With so many factors influencing this data set, one should question whether it is even worthy of exploration and publication. To that I say yes, so long as the reader interprets the conclusions with the above factors in mind. It is my belief these initial steps towards large-scale giant pumpkin genetic performance quantification are an improvement beyond the current, highly subjective method of seed selection. Use this data as a tool in your process rather than a definitive guide for seed selection. There. Having hopefully covered my butt from critical purist backlash, here's the list of top performing seeds based on the average of their progeny:

Rank	Seed	n=	ave lbs	Rank	Seed	n	ave lbs
1	2096 Meier	13	1504	19	1892 Faust	13	1173
2	1781 Zywiec	10	1504	20	2008 Neptune	30	1172
3	2624 Willemijns	26	1455	21	1985 Miller	22	1159
4	2145 McMullen	154	1450	22	2307 Wallace	10	1131
5	2009 Wallace	17	1412	23	1991.5 Vander Wielen	10	1121
6	2230 Wallace	46	1340	24	2185 Brandt	8	1117
7	1756 Howell/Jolivette	21	1320	25	2020.5 Werner	17	1105
8	2090.5 Stelts	9	1295	26	1916 barron	23	1096
9	2066 Geddes	19	1281	27	2269 Paton	26	1070
10	1803 Hoelke	22	1268	28	1998 Jutras	21	1062
11	2075 Connolly	13	1250	29	1790 Wallace	30	1050
12	2043 Clementz	11	1238	30	2109 Zywiec	15	1044
13	1824 Shenoha	14	1238	31	2095 Willemijns	28	1032
14	2095 Brandt	11	1231	32	2059 Daletas	10	991
15	2363 Holland	55	1212	33	2017 Martin	16	970
16	1937 Urena	33	1211	34	2002 Tobeck	20	914
17	2261 Wallace	27	1199	35	2323 Meier	8	819
18	2004 Vander Wielen	15	1194	36	2036 Glasier	10	804

The data here suggest that the 2096 Meier and 1781 Zywiec seeds have produced the highest average weight progeny for all seeds that have 8 or more official progeny entries in the database. Note that these seeds have 13 and 10 official entries, respectively, and that the statistical significance of these outcomes may not be there due to lack of volume. The 2145 McMullen seed has the most progeny (154 entries) of any seed in the database and ranks 4th in this list. Is this seed such an outstanding performer because it truly has the highest genetic upside or is it because the growers who planted this seed, from heavy hitters to average Joe's, gave it the most attention in their patches, which led to great outcomes, which further perpetuated this seed's popularity and subsequent above-average performance? That is the primary flaw in interpreting these results, so once again, proceed with caution. The average progeny weight of these 35 top-end seeds is 1179lbs. With that, you can calculate which seeds are cream of the crop by leaning towards those with progeny average above this threshold.

Let's transition to an analysis that I feel carries more statistical significance. I decided to break maternal lineage of the 5324 official GPC entries into 100lb increments, starting from 1600-1699lbs seed/mother and categorizing all the way up to 2600-2699lbs. I then averaged all the entries/progeny within each of those 100lb seed categories to determine if average weights demonstrated any notable trends as maternal (seed) weight increased. As alluded to earlier in this discussion, the data here imply that growing seeds from the biggest pumpkins increases your chances of a bigger outcome in the patch (surprise surprise):



A couple caveats regarding the graph above: The dataset implies that seeds in the 2100-2199lb range produce the largest progeny. This is obviously skewed by the overwhelming number of data points from the 2145 McMullen. The true skew in the data results from a limited number of maternal lines & progeny above the 2199lb threshold. Only 5 maternal lines populate the 2200-2299lb category and only 3 populate the 2300-2399lb category. Obviously there's only one genetic line above 2600lbs. With a narrow band of maternal lines, coupled with a significantly lower number of official progeny produced from these lines relative to the 2100lb category, their collective averages are lower.

At this point, it should be asked why the 2600lb category demonstrates the highest average progeny weight. The 1414lb average weight from a single seed, the WR 2624 Willemijns, is populated by 26 official GPC entries. What can we draw from this? Is the 2624 a "silver bullet" seed that locks in maximum upside in seed genetics? 3-way cross weight average analysis $(2624+2145F+1872M)/3=2214$ lbs places it near the top of the "biggest" crosses available in the hobby. For reference, as of 2018, the largest 3-way CWA appears to be Ron Wallace's 2114 $(2145 \times 2624) = 2294$ CWA. (email me at joeailts@gmail.com if you find a larger CWA pumpkin). It is impossible to claim that the largest pumpkin in the world has produced the highest upside seeds in the hobby. But the narrative fits, doesn't it? Time will tell if this proves to be true as the next season rolls around and more observations can be collected. In the interim, I believe we can look at this dataset as a whole and reasonably conclude that seeds from the largest pumpkins provide the likeliest chances of producing a larger pumpkin. The R-squared value of the linear fit to the data is 0.7308. A value of 1.0 is a perfect fit, so we're in the ballpark and it tells us that the upward trending line across the chart can be reasonably assumed to be true.

In summary, the analysis herein is a first-look at quantifying the seed genetic selection process. While crude and lacking true statistical significance in a number of the observations made, it does serve as a starting point when considering next year's lineup. We can probably assume that planting seeds from the largest pumpkins increases our chances of growing a larger pumpkin. That seems like a "duh"

statement in hindsight, but it is reassuring to put some numbers behind that assumption. And therein lies the goal...refine the process to maximize upside, using all the tools at our disposal, in the interest of growing the largest pumpkins we possibly can.

Best of luck in the patches and your seed selection process!

-Joe